

**UNIVERSIDAD COMPLUTENSE DE MADRID**

**FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES**



**PREDICCIÓN DE CRISIS EMPRESARIALES EN SEGUROS NO  
VIDA MEDIANTE LA METODOLOGÍA ROUGH SET**

**MEMORIA PARA OPTAR AL GRADO DE DOCTOR**

**PRESENTADA POR**

María Jesús Segovia Vargas

Bajo la dirección de los doctores

José Antonio Gil Fana

Antonio Heras Martínez

**Madrid, 2003**

**ISBN: 84-669-2278-4**

**FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES  
UNIVERSIDAD COMPLUTENSE DE MADRID**

**TESIS DOCTORAL**

**PREDICCIÓN DE CRISIS EMPRESARIALES EN  
SEGUROS NO VIDA MEDIANTE LA METODOLOGÍA  
ROUGH SET**

**AUTORA: María Jesús Segovia Vargas**

**DIRECTORES: José Antonio Gil Fana**

**Antonio Heras Martínez**

**Enero 2003**

**-INDICE-**

<b>INTRODUCCIÓN.....</b>	<b>1</b>
<b>CAPÍTULO I.- EL MÉTODO ROUGH SET.....</b>	<b>22</b>
I.- Introducción: análisis de la decisión, modelo de preferencias y enfoque rough set.....	23
II.- Aproximación al concepto de rough set.....	34
III.- Tabla de información y relación de no diferenciación.....	37
IV.- Aproximación de conjuntos.....	43
V.- Aproximación de la partición de $U$ o clasificación rough.....	47
VI.- Reducción y dependencia de atributos.....	49
VII.- Reglas de decisión.....	54
VIII.- Ayuda a la toma de decisiones utilizando reglas de decisión.....	68
IX.- Tipos de problemas de decisión.....	82
<b>CAPÍTULO II.- APLICACIONES DEL MÉTODO ROUGH SET.....</b>	<b>95</b>
I.- Introducción.....	96
II.- La predicción de las crisis empresariales.....	106

**CAPÍTULO III.- SISTEMA ROSE: EXPLORADOR DE DATOS****ROUGH SET .....122**

I.- Antecedentes del ROSE y otros algoritmos existentes.....123

II.- ROSE – Software de aplicación de la Teoría Rough Set.....130

**CAPÍTULO IV.- LAS VARIABLES DEL MODELO.....144**I.- La variable dependiente: concepto de crisis empresarial.  
Insolvencia.....145II.- Las variables dependientes: los estados financieros como  
instrumentos de la predicción de las crisis empresariales.  
Los ratios.....151II.1.- Introducción: enfoque metodológico seguido en  
la selección de los ratios.....151

II.2.- Selección de la muestra.....160

II.3.- Análisis del Balance y la Cuenta de Resultados  
de las entidades aseguradoras.....166

II.4.- Elección de las variables: Los Ratios Contables.....174

ANEXO CAPÍTULO 4.....195

**CAPÍTULO V.- EL MÉTODO ROUGH SET APLICADO****A LA PREDICCIÓN DEL FRACASO****EMPRESARIAL EN LAS EMPRESAS****DE SEGUROS NO VIDA EN EL CASO****ESPAÑOL.....198**

I.- Fundamentación del Análisis.....199

II.- Aplicación del Análisis y Resultados.....201

II.1.- Definición del Problema.....201

II.2.- Aplicación de la metodología Rough Set.....	202
II.3.- Resultados.....	227
III.- Validación de los resultados: Clasificación de empresas utilizando los dos algoritmos generados.....	231
IV.- Comparación del Análisis Rough Set con el Análisis Discriminante.....	233
ANEXO CAPÍTULO 5.....	240
<b>CONCLUSIONES.....</b>	<b>246</b>
<b>BIBLIOGRAFÍA.....</b>	<b>258</b>

# **INTRODUCCIÓN**

Todas las decisiones financieras de una institución o corporación (un banco, una empresa, una compañía de seguros, etc...) implican, en mayor o menor medida, la idea de optimización. Esta idea subyace en los problemas de decisión en el ámbito financiero, tanto para el largo como para el corto plazo: búsqueda de la estructura de capital más adecuada, la oportuna colocación de fondos, la gestión del capital circulante, inventarios, tesorería, deudas a corto plazo, deudas a cobrar, etc., además de los análisis de inversiones.

Además, muchas decisiones financieras implican la clasificación de una observación (empresa, título de una cartera, ...) en una categoría o en grupo. Un caso especial de este tipo de problemas aparece cuando la clasificación es binaria, es decir cuando el número de grupos se limita a dos. Existe una gran cantidad de literatura financiera que se ha dedicado a este tipo de problemas de clasificación en dos grupos, entre los cuales destacamos: problemas de clasificación de créditos entre fallidos y no, fusiones y adquisiciones, clasificación de bonos, ..., o la predicción del fracaso empresarial, que será el objetivo fundamental de esta tesis.

Centrándonos en este último problema cabría preguntarse sobre ***por qué predecir las crisis empresariales.***

La viabilidad y, en consecuencia, el riesgo de fracaso empresarial ha sido reconocido como un factor importante en el área de la macroeconomía o en los análisis industriales. Al mismo tiempo, tanto el número de empresas quebradas como la tendencia relativa, son considerados indicadores de la solidez de las industrias y del conjunto de la economía. De hecho, la importancia del riesgo del fracaso ha sido un factor tenido en cuenta desde hace mucho tiempo, y esta es la razón que explica el gran desarrollo de las investigaciones en el área de la gestión financiera para la evaluación de las empresas y la declaración de su viabilidad.

El fracaso afecta a totalidad de la empresa y tiene un alto coste (en términos de efectos perniciosos) tanto para la propia empresa, para sus colaboradores, para la sociedad y para la totalidad de la economía de un país. Además, el entorno económico actual (con situaciones coyunturales de crisis económica) en el que las empresas están obligadas a desenvolverse, hace necesario anticipar el conocimiento de los posibles problemas de falta de rentabilidad o estabilidad de la misma.

En consecuencia, la predicción de la insolvencia empresarial es una cuestión de vital importancia para todas aquellas partes relacionadas con una empresa: propietarios o accionistas, directivos, trabajadores, prestamistas, proveedores, clientes, la sociedad y las autoridades. Necesitan conocer, con el mayor grado de exactitud posible, el nivel de solvencia de la empresa para poder establecer con suficiente tiempo de antelación políticas de ajuste que minimicen el riesgo y las negativas consecuencias de posibles situaciones de crisis empresarial no deseadas.

Por otro lado, la teoría financiera analiza sus problemas de decisión (no solo la insolvencia empresarial, sino también los mencionados al comienzo de esta introducción), tanto a corto como a largo plazo, según una perspectiva de búsqueda de soluciones óptimas considerando un único objetivo y un conjunto de soluciones posibles. Este hecho, ha propiciado que muchos investigadores hayan aplicado métodos de investigación operativa a los problemas financieros. Esta perspectiva es demasiado simplista y dificulta su aplicación a complejos problemas financieros reales.

Sin embargo, a partir de la década de los setenta, el análisis de estos problemas financieros se empieza a producir en un marco más realista basado en el paradigma de la decisión multicriterio. Este nuevo escenario se fundamenta sobre la base de que los decisores, en el ámbito económico, buscan un equilibrio entre una serie de objetivos que normalmente están en conflicto, teniendo en cuenta todo un conjunto de criterios que reflejan sus preferencias, su conocimiento y experiencia. Es



decir, se pasa de un contexto de búsqueda de soluciones óptimas a otro que busca soluciones satisfactorias.

El principal objetivo del paradigma multicriterio es proveer a los decisores de herramientas que les permitan resolver, sobre la base de múltiples criterios, complejos problemas de decisión donde hay varios objetivos en conflicto.

Hemos, en consecuencia con lo expuesto, acudido a la búsqueda de un método multicriterio que se ajuste a la investigación que vamos a realizar.

Por tanto, **la elección de un método multicriterio se justifica** por las ventajas que presenta su utilización así como por la problemática que plantea la utilización de otros métodos, como pueden ser los estadísticos.

En cuanto a las *ventajas* que presentan los métodos multicriterio cabría mencionar la posibilidad de estructurar problemas de decisión complejos al permitir modelizarlos incorporando simultáneamente varios objetivos y varios criterios.

Otra ventaja es que permiten la utilización tanto de variables cuantitativas (ratios financieros) como variables de tipo cualitativo. Los problemas reales obedecen a un conjunto de causas complejas y múltiples que justifican la necesidad de incorporar ambos tipos de variables para poder analizarlos. De este modo, los criterios cualitativos tales como la capacidad de la dirección, la tendencia del mercado, la posición en el mercado, etc., juegan un papel fundamental en el proceso de decisión, incluso en algunos casos las variables cuantitativas son consecuencia de algunos factores cualitativos. Por tanto, y dada la complejidad del problema que estamos considerando, los métodos multicriterio se adaptan perfectamente a él.

Además, estos métodos permiten que el usuario final, el decisor, se pueda implicar en todo el proceso de toma de decisiones incorporando su

modo de pensar, conocimiento y experiencia en forma de criterios y evaluando los resultados que se obtienen lo que le permite utilizar dichos resultados para justificar las decisiones que van a tomar. En consecuencia, las decisiones que proponen los modelos obtenidos con estos métodos, tienen que estar de acuerdo con las decisiones de los centros decisores que son sus naturales usuarios. Es más, los decisores deberían ser capaces de comprender y explicar las decisiones y, por esta razón en la elaboración de los modelos multicriterio se utiliza su conocimiento y experiencia. En definitiva, este enfoque permite a los centros decisores comprender el funcionamiento del modelo y enseñarlos para que puedan incluir nuevos conocimientos y experiencia. De esta manera se pueden modificar los modelos para hacer frente a los cambios del medio y explorar otros nuevos caminos.

En cuanto a la *problemática* que plantea la utilización de otras técnicas, como pueden ser las de tipo estadístico y otras utilizadas en el pasado, es que no siempre han sido capaces de dar suficiente respuesta a los problemas prácticos a los que se enfrentan los usuarios. Un problema que plantean estas técnicas es el hecho de que el cumplimiento de las hipótesis estadísticas que subyacen a las mismas no siempre es fácil (por ejemplo, los ratios financieros son variables muy utilizadas en la predicción de crisis empresariales y la mayoría de ellos no se ajustan a una distribución normal, lo cual puede, en principio, cuestionar sus resultados o complicar el análisis). Otro problema es la dificultad en la explicación de los resultados de los modelos por los usuarios ya que la mayoría de los modelos emiten sus juicios sin una racionalización adicional.

Por estas y otras razones, muchos investigadores y usuarios han encontrado los métodos multicriterio interesantes y atractivos y los han utilizados en sus investigaciones.

Dentro de los posibles métodos multicriterio, las herramientas surgidas de la *Inteligencia Artificial* se han mostrado muy apropiadas para

tratar problemas financieros. La filosofía de la que parten es diferente, concretamente pretenden ayudar a la toma de decisiones pero simplificando la tarea del usuario final, de tal forma que a este último no se le requiera prácticamente ningún conocimiento técnico para interpretar los resultados del modelo o para su aplicación.

La *Teoría Rough Set* (Pawlak, 1982, 1984, 1991) es un enfoque que se encuadra dentro de las aplicaciones de la Inteligencia Artificial. Utiliza la experiencia de una forma objetiva, a través del estudio de la experiencia histórica de manera cuantitativa, para así explicitar reglas que, resumiendo y objetivando esa experiencia acumulada, ayuden en las decisiones futuras.

El problema que nos ocupa, la predicción de las crisis empresariales, no permite establecer fácilmente las relaciones causa-efecto entre los atributos que están relacionados o que pueden llevar a una quiebra empresarial y el acaecimiento efectivo de la misma, por varias razones (McKee, 2000). Una de ellas es que en la actualidad la teoría no nos permite identificar claramente todos los posibles atributos relevantes, y, en consecuencia, algunos podrían ser omitidos. Una segunda razón es que algunos de los atributos pueden ser cualitativos y otros pueden ser cuantitativos, lo que provoca un problema adicional de medida. Una tercera razón es que los atributos pueden ocurrir en uno o más periodos previos a la quiebra y de este modo surge el problema de la medición temporal. Cuando se trabaja con compañías reales, estas dificultades se traducen en que el análisis de las relaciones causa-efecto se compone normalmente de inconsistencias en las clasificaciones. Es decir, una empresa sana puede tener los mismos atributos que una quebrada. Un enfoque para tratar esta problemática es el Rough Set.

Esta metodología presenta, entre otras, las siguientes ventajas:

- a) Con este análisis se obtiene como resultado la reducción de un sistema de información completo y compuesto por un gran

número de casos consiguiéndose un modelo reducido que consiste en una descripción generalizada de ese conocimiento (sistema de información) que estamos considerando.

- b) El modelo obtenido se compone de un conjunto de reglas de decisión fácilmente entendibles por lo que normalmente no necesitan interpretación.
- c) Cada regla de decisión se justifica por un conjunto de ejemplos reales, por tanto, pueden usarse para aplicaciones prácticas del mundo real.
- d) No es necesaria ninguna información adicional, como probabilidades (en la estadística) o grado de pertenencia (en la teoría fuzzy set).

Todas estas ventajas, y las que comprobaremos a lo largo de nuestra exposición justifican la elección de esta metodología pero existe otra razón adicional.

La utilización de un método multicriterio para abordar decisiones financieras en tiempo real, implica la necesidad de disponer de un software que permita la aplicación práctica de dicho método. La metodología Rough Set ha sido implantada en varios software lo cual permite su uso para acometer problemas reales. Queremos mencionar que los trabajos empíricos publicados en el ámbito económico y financiero que utilizan esta metodología se han efectuado con el software RoughDas antecesor de ROSE, que es el que nos propusimos utilizar en la parte empírica de nuestra investigación. Por tanto, hasta la fecha, en el ámbito económico-financiero no conocemos ningún trabajo en el que se emplee este nuevo software para abordar el problema financiero real de la predicción del fracaso empresarial. Nuestro objetivo será ver si es viable su utilización para acometer este tipo de problemas.

El software utilizado es el programa ROSE facilitado por la Universidad de Poznan (Polonia) y desde aquí queremos expresar nuestro agradecimiento tanto al *Institute of Computing Science* de la Universidad de Poznan como al Profesor Wilk por su inestimable colaboración. Queremos mencionar que cualquier ordenador personal conectado a internet puede acceder a la página web [www-idss.cs.put.poznan.pl/rose](http://www-idss.cs.put.poznan.pl/rose). donde se puede descargar el programa ROSE, su manual e informarse brevemente de algunas de sus características.

Justificada la elección del problema financiero que vamos a abordar y la metodología que vamos a emplear podríamos concretar **un primer objetivo** de nuestro trabajo que es el desarrollar **un modelo**.

Los modelos son una herramienta más para los usuarios, que apoyados en otras informaciones, pueden ayudarles a dictaminar sobre una determinada empresa como empresa en crisis.

Es decir, un modelo fiable constituye un gran valor, en primer lugar para los empresarios o gerentes, ya que lo pueden usar para identificar y controlar los factores críticos que pueden influir en un aumento de la probabilidad de que la empresa fracase (estaríamos ante un uso *activo* de modelo). En segundo lugar el modelo puede ser usado tanto por analistas, autoridades, empresas de capital riesgo, inversores, trabajadores o sus representantes..., en sus tomas de decisiones a la hora de evaluar el riesgo financiero asociado a la empresa que estemos considerando (estaríamos ante un uso *pasivo* del modelo) (Laitinen, 1992).

Queremos señalar, que los modelos no pretenden sustituir la capacidad o intuición de los distintos entes interesados en conocer la evolución de la empresa, pero sí van a proporcionar información adicional sobre la solvencia de la misma. Es decir, complementarían la capacidad de juicio y experiencia de las personas, y añadirían una sistemática o procedimientos objetivos comunes de análisis.

Es decir, pretendemos proveer de una herramienta útil para la toma de decisiones de los distintos usuarios de la información contable.

Este modelo, en términos prácticos, puede usarse como “sistema de alerta temprana” para directivos, autoridades y demás usuarios interesados en prevenir la crisis empresarial (Dimitras et al., 1995). Es decir, los analistas y otros usuarios podrían establecer con suficiente tiempo de antelación políticas que minimicen el riesgo de posibles situaciones de crisis empresarial no deseadas. Otra utilidad sería la utilización del modelo como sistema de diagnóstico automático para preseleccionar aquellas empresas que necesiten una atención especial, de una manera rápida y a un coste bajo, y de este modo se lograría una gestión más eficiente del tiempo dedicado por el analista financiero, auditor, autoridad supervisora, etc. Finalmente, otra aplicación del mismo podría ser para dar uniformidad a los juicios y métodos efectuados sobre una empresa.

Cualquier modelo necesita la especificación de las **variables** que van a formar parte del mismo. El problema de predicción del fracaso empresarial es uno de los problemas más complejos dentro del área de gestión empresarial ya que obedece a multitud de causas que pueden únicamente describirse por un gran número de variables.

Los ratios financieros han sido las principales variables utilizadas en relación a la predicción de crisis empresariales. Aunque la metodología Rough Set permite la utilización de variables cualitativas (además de cuantitativas) como pueden ser la experiencia de los directivos, el nicho de mercado en el cual se encuentra la empresa, el entorno y sus cambios, etc., no hemos considerado en nuestra investigación características cualitativas debido a la imposibilidad de conseguirlas.

Por ello hemos recurrido a la utilización de documentos contables. Éstos nos permiten obtener una información cuantificable, objetiva y sistemática de la realidad económica y financiera de la empresa. Y, a

través de su análisis, mediante determinadas técnicas eficientes, se puede evaluar e interpretar la información contable disponible por el analista, para ofrecer un diagnóstico sobre la situación pasada, presente o futura. Es más, los modelos de predicción del fracaso empresarial proporcionan una forma de probar el contenido informativo de los datos contables sobre la solvencia de la empresa, al medir la relación entre los datos y el fracaso. A la vez, son un mecanismo para pronosticar el fracaso empresarial, de manera que proveen una herramienta útil para la toma de decisiones por parte de los múltiples usuarios del análisis de la solvencia (Calvo-Flores y García, 1998a, p. 8).

En este sentido, los ratios financieros individualmente considerados ofrecen una valiosa información sobre aspectos diferentes de la situación empresarial y además, pueden agruparse en estructuras de análisis con el fin de sumar la información que cada ratio contiene sobre la empresa. De esta forma, conoceremos la valoración global de la situación empresarial y los aspectos concretos a mejorar para superar el deterioro global, en caso de que dicha valoración advierta del mismo, o en cualquier caso para corregir desequilibrios. De hecho, muchos de los trabajos realizados en relación con la predicción del fracaso empresarial han demostrado que existe una elevada correlación entre los datos contables y la futura solvencia de la empresa, corroborando la utilidad de dichos datos para la toma de decisiones.

Es más, el análisis contable es una disciplina viva y en constante evolución que se nutre de diferentes herramientas para cumplir el objetivo de convertir los datos contenidos en los estados económico-financieros en información útil para la toma de decisiones. Esta disciplina ha ido evolucionando en cuanto a la utilización de los instrumentos técnicos que emplea paralelamente a la evolución de otras ciencias, como la Estadística, la Investigación Operativa y la Computación. El desarrollo de nuevas técnicas de computación procedentes de la Inteligencia Artificial

ha revolucionado el análisis de estados financieros y se revela como un área en constante crecimiento (Serrano y Martín del Brío, 1993).

Pero estamos ante instrumentos de carácter más empírico que teórico. Es decir, no existe una teoría elaborada que investigue las causas de quiebra o fracaso a través de la evolución de los ratios contables. Falta el sustrato teórico que sirva de nexo de unión entre los diversos métodos y modelos.

Es decir, en los trabajos empíricos sobre la predicción de la quiebra no se parte de una teoría porque no existe, en sentido riguroso, una teoría económica sobre la solvencia empresarial. La acumulación de trabajos empíricos permite plantear hipótesis sobre qué variables contables predicen la futura situación financiera de la empresa, lo cual puede ser un pilar a la hora de elaborar una teoría contable sobre la solvencia empresarial en un futuro.

Por tanto, sabemos muchas cosas acerca de cómo las empresas desembocan en un fracaso económico, sobre cómo es el proceso de crisis y el deterioro de las compañías, pero no tenemos una teoría completa.

Con este trabajo pretendemos, como **segundo objetivo**, contribuir al desarrollo de la base que permita abordar convenientemente la teoría financiera de la insolvencia y desarrollar investigaciones empíricas que permitan el desarrollo de modelos generalmente aceptados. En este sentido, los métodos multicriterio suministran el medio para las aplicaciones prácticas y pueden ayudar a la realización de ese sustrato necesario para el análisis del problema del fracaso empresarial que permita desarrollar una teoría sobre la predicción del mismo.

En este punto, es necesario mencionar que la mayoría de estudios y trabajos llevados a cabo para predecir la insolvencia o el fracaso empresarial no han considerado **el sector** seguros. Este hecho se explica



en parte por las siguientes razones: una equivocada falsa importancia debido al escaso número de empresas que componen este sector en comparación con otros y, en segundo lugar, las peculiaridades del negocio asegurador llevan a la necesidad de poseer unos conocimientos técnicos específicos sobre el mismo que lleva a muchos investigadores a no acometer su análisis.

Estas peculiaridades hacen que variables y modelos que son válidos en otros sectores no lo sean tanto en el sector asegurador o deban ser redefinidos y adaptados a estas especiales características que presenta el mismo.

Es necesario mencionar que hasta la fecha se han elaborado pocos modelos de predicción de insolvencias en empresas de seguros. La mayoría de los escasos trabajos utilizan métodos estadísticos con los problemas que ya hemos mencionado. Por tanto, aplicar la metodología Rough Set podría convertirse en una herramienta adicional, a la par que novedosa, para este sector. En consecuencia, nos centraremos en el sector de seguros.

La insolvencia, su temprana detección o el conocimiento de las condiciones que pueden conducir a ella, en una compañía de seguros es una de las principales preocupaciones de legisladores, consumidores y directivos de este tipo de entidades. Esta preocupación surge como resultado de la necesidad de proteger al público de las consecuencias de las insolvencias de los aseguradores, por un lado, y la necesidad de minimizar la carga que supone para el estado hacer frente a las mismas a través de los fondos de garantía, por otro.

Además una función importante de los gobiernos es regular el sector asegurador y a través de ella, controlar la solvencia del mismo, entre otras razones porque dada la naturaleza de este tipo de negocio las primas se pagan de manera previa a que ocurra el pago del eventual siniestro cubierto por ella. Por tanto el control por parte de las autoridades

de este sector reduce la posibilidad de que los futuros siniestros no sean pagados.

En la mayoría de los países desarrollados se han tomado medidas para prevenir crisis financieras en las entidades aseguradoras y evitar las repercusiones negativas que tendrían sobre los colectivos afectados. Se pretende, en definitiva, el control de la actividad aseguradora para lo cual se han articulado diversos sistemas de control: obligación de informar sobre su situación económico-financiera mediante la publicidad de las cuentas anuales y obligándolas a ser auditadas, imposición de normas encaminadas a garantizar su solvencia como pueden ser el cumplimiento de los coeficientes de garantía y de solvencia, provisiones obligatorias, inversión de las mismas, capital mínimo, regulación de los precios, remisión a la autoridad supervisora de los estados confidenciales, o, finalmente, realización de inspecciones in situ.

En definitiva, el análisis de la solvencia en cualquier sector es importante, pero la creciente importancia de las entidades aseguradoras en cualquier sistema financiero, hace que dicha cuestión sea clave en este tipo de entidades.

Las características propias del sector seguros motivan el preguntarnos qué factores hay que tener en cuenta a la hora de analizar la insolvencia, y como consecuencia de ella, el fracaso y las crisis dentro del mismo.

Para analizar la solvencia de las compañías aseguradoras hay que tener en cuenta multitud de *factores*, los comunes a cualquier sector a los que habría que añadir los específicos del sector asegurador.

Vamos a extendernos en esta cuestión, ya que al habernos centrado en los ratios financieros por los motivos argumentados, el resto de factores que se podrían contemplar caen fuera de nuestro estudio.

Siguiendo a Bannister (1997), un primer conjunto de factores serían *los factores macroeconómicos*. Estos son los que vienen determinados porque la compañía está operando en un determinado país. Dentro de estos tendríamos.

- La práctica contable. Las compañías de seguros están sometidas a dos tipos de normas en este terreno las aplicadas a todas las compañías en general y las específicas para el sector. La razón de estas últimas es la protección de los asegurados, lo cual lleva a un incremento de exigencias de información que tienen su impacto en la contabilidad.
- La supervisión a que está sometido el sector: Es necesario revisar si la compañía en cuestión está cumpliendo con todos los requisitos y la legislación aplicable al mundo del seguro.
- La práctica fiscal: La fiscalidad de un país tiene una enorme relevancia en las operaciones de seguro. Una fiscalidad favorable (la existencia de incentivos fiscales) a la actividad aseguradora puede llevar a un aumento de los fondos procedentes de suscripciones de primas y viceversa.
- Factores monetarios. Las fluctuaciones en el tipo de cambio de las monedas pueden influir en la industria aseguradora internacional ya que puede haber un aumento en la siniestralidad derivada de una diferencia de cambio negativa. Es necesario además revisar la convertibilidad de las reservas o del efectivo cuando los pagos se han de hacer en monedas diferentes a la nacional.
- Factores económicos y políticos. Una amplia gama de factores económicos y políticos pueden afectar a la capacidad de una aseguradora para hacer frente al pago de los siniestros; entre ellos podemos distinguir: la inflación y su impacto en los

resultados o en los fondos, la situación bursátil que puede afectar al valor de las inversiones y provisiones, cambios legales que lleven como consecuencia un aumento de las reclamaciones.

- El conjunto de la industria aseguradora de un determinado país. Es necesario analizar la competitividad dentro del sector, su capacidad financiera.

Un segundo grupo de factores serían los denominados *factores corporativos*, es decir todos aquellos que pueden afectar a un asegurador o reasegurador considerado de manera individual. En otras palabras, aquellos que tratarían de medir la calidad de un asegurador o reasegurador teniendo en cuenta factores cuantitativos (ratios, etc.) y cualitativos (todos aquellos factores no financieros que pueden afectar a la capacidad o voluntad de pagar). En este segundo grupo destacaríamos:

- La estructura de grupo al que pertenece la compañía de seguros que estemos considerando. Desde el punto de vista del análisis de la solvencia la propiedad puede dividirse en tres grupos:
  - a) Grupo asegurador cotizado en bolsa donde el negocio asegurador es la principal actividad y la propiedad no está concentrada. En este caso habrá que centrarse no tanto en los cambios de la propiedad sino en los directores y gestores, sus cambios y su trayectoria.
  - b) Grupo privado asegurador. Aquí al no estar tan “diluida” la propiedad sí es necesario ver los cambios en la misma y la presión que la propiedad ejerce sobre los directivos y ejecutivos, así como su conocimiento del sector asegurador.

- c) Grupo no asegurador ya sea cotizado en bolsa o privado. Aquí es especialmente importante desde el punto de vista de la solvencia analizar la marcha del negocio principal, ya que si éste sufre una recesión, no estará en disposición de respaldar a la compañía aseguradora (si esta está en crisis) por no ser la actividad aseguradora su actividad principal.
- Los directivos. El equipo directivo de una empresa aseguradora o reaseguradora varía enormemente en cuanto a experiencia, competencia y grado de responsabilidad. Cualquier analista debe prestar atención a la composición del equipo directivo, sus cambios, su trayectoria en la compañía, el tiempo que llevan en la misma o dentro del sector, etc. Es, además, necesario comprobar cuál es la capacidad del equipo directivo en relación a los principales competidores.
- La imagen. Cada asegurador tiene una imagen frente a sus clientes, reaseguradores y con el mundo del seguro en general. La imagen que se tenga de una compañía puede ser un factor positivo a la hora de valorar la solvencia o no. Lo importante es considerar si esta variable es relevante para una compañía a la hora de analizar su solvencia y qué factores pueden dañar seriamente dicha imagen.
- La estrategia de mercado. Al igual que ocurre en otros mercados, en éste cada compañía compite por su cuota de mercado y su éxito en esta estrategia depende de factores que incluyen cualquier mínima ventaja en su habilidad para hacer frente a los siniestros, su relación con los clientes y la tenencia de cuentas históricas, la efectividad de un programa de reaseguro, un buen marketing, etc. Todas estas cuestiones son interesantes de tener en consideración de cara a evaluar futuros resultados financieros y la probabilidad de que estos realmente acaezcan, es decir, una mala política de negocio es un

indicador de que en cualquier momento los resultados pueden caer y afectar al pago de siniestros y en consecuencia, a la solvencia.

- Los agentes de seguros. Es necesario un adecuado control tanto de que los resultados que producen como de la imagen que dan de la compañía.

Finalmente, el tercer grupo de factores que tenemos que tener en cuenta son los *factores específicos* de la compañía. Dentro de esta categoría estarían

- Los fondos propios. Desde el punto de vista de la solvencia, es necesario que los mismos sean adecuados ya que estos fondos pueden ser exigidos ante una eventualidad derivada de una mala tarificación, depreciación del capital o una mala política reaseguradora.
- La rentabilidad. Existen tres factores que ejercen una gran presión sobre la solvencia y sobre los fondos propios y que pueden ser medidos en términos financieros utilizando ratios.
  - a) Rentabilidad del negocio. Se puede medir a través del ratio combinado.
  - b) Rentabilidad de las inversiones. Existen varios ratios para medir la rentabilidad de las inversiones y que son susceptibles de ser, en consecuencia, utilizados. Por citar algunos de ellos: ingresos financieros sobre activo de inversiones, ingresos financieros sobre gastos técnicos directos e ingresos financieros sobre primas y recargos del seguro directo.
  - c) La depreciación o revalorización del capital. Se puede medir a través del cambio en el valor de las inversiones (valor de

mercado normalmente) sobre el valor original de las mismas (es decir, el valor al principio del periodo que estamos considerando).

- La viabilidad financiera. La continuidad con éxito de las operaciones financieras exige la generación de beneficios suficientes para satisfacer (después del pago de impuestos) las expectativas de los accionistas en forma de dividendos y obtener suficientes reservas para aumentar el capital base. Como mínimo debería generarse suficiente beneficio para mantener el capital base en un crecimiento al menos igual a la inflación. Por tanto, es particularmente importante desde el punto de vista de la solvencia el análisis de la tendencia de los resultados, ya que cualquier caída o variación en su continuidad es una señal de peligro.
- El margen de solvencia. El margen de solvencia obligatorio normalmente toma un valor arbitrario y de acuerdo a una fórmula legislada. El analista debe recalcular ese margen y comprobar si al menos cumple con los requisitos mínimos legales o por el contrario excede de dichos requisitos.
- El reaseguro. En cualquier análisis de la solvencia de una compañía de seguros una variable clave es el reaseguro.

Este tercer grupo de factores a tener en cuenta, como vemos se pueden medir directamente a través de ratios financieros proporcionando así un marco para el análisis financiero. Es muy importante de cara a cualquier análisis financiero el estar familiarizado tanto con los ratios como con los factores que subyacen bajo ellos y que les afectan. La comparación de los cambios anuales en estos ratios ayudará a determinar la tendencia y a través de esta determinar la proyección de la empresa. Cuando se analizan los valores de varios años de los ratios es necesario comprobar la tendencia de las pérdidas y ganancias del negocio

asegurador, la rentabilidad de las inversiones, y el riesgo de tipo de cambio.

Una vez que se han considerado los factores que podríamos considerar relevantes para el análisis de la solvencia de una entidad aseguradora, éstos deberían ser incorporados a cualquier modelo que queramos utilizar para predecir el fracaso empresarial en forma de variables ya sean de tipo cualitativo o de tipo cuantitativo.

Hemos de recalcar que la imposibilidad de recabar información sobre los otros dos grupos de factores han motivado que en nuestro trabajo solo incluyamos el último grupo, sin embargo dado que el **tercer objetivo** es mostrar la idoneidad del análisis Rough Set para la predicción de crisis empresariales en el sector asegurador, no sería difícil (ya que esta metodología permite la incorporación de variables de tipo cualitativo) incorporar, si se dispone de información, cualquier factor que consideremos relevante, y de esta forma mejorar el modelo que obtengamos.

Para cumplir este requisito tenemos que ver cuales son las *principales fuentes de información de las compañías aseguradoras*. Podríamos enumerar las siguientes: a) *Los datos suministrados por la propia compañía*, particularmente sus cuentas anuales y los estados complementarios que las acompañan (incluyendo los estados financieros de carácter interno); b) *Los informes remitidos por la compañía aseguradora a la autoridad supervisora*. Normalmente no son públicos pero si se dispone de ellos porque los facilite la compañía presentan una gran ventaja para cualquier análisis ya que presentan una mejor consistencia y preparación, e incluso un mayor detalle que las cuentas presentadas y publicadas de la compañía; c) *Otros informes oficiales*, principalmente los datos suministrados a las rectoras de las bolsas; d) *Cuestionarios o charlas informales y formales con la compañía*; e) *Material publicado*, principalmente comentarios sobre la compañía y el mercado en el que opera tanto en revistas del sector como en la prensa



general y financiera; f) *Datos sobre la solvencia realizados por especialistas*, mediante servicios de suscripción (por ejemplo, Standard & Poor, etc.); g) *Análisis de los corredores de bolsa, de los departamentos de inversiones de las compañías aseguradoras, etc.*

Sin embargo, en nuestro análisis no hemos dispuesto de la gran mayoría de las fuentes de información que hemos mencionado, por ello, nuestras fuentes de información han sido los balances y las cuentas de pérdidas y ganancias de las empresas, y las variables que hemos considerado relevantes de cara al análisis han sido los ratios.

*En resumen, a lo largo de la introducción hemos intentado responder a las tres cuestiones que nos hemos planteado: por qué predecir crisis empresariales; por qué aplicar un método multicriterio y, en concreto, la teoría Rough Set; y, por qué aplicarlo al sector asegurador.*

*En consecuencia, nuestro trabajo trata de comprobar empíricamente la utilidad e idoneidad del método Rough Set para el análisis de la solvencia de las empresas de seguros, mediante el desarrollo de un modelo que ha sido aplicado a empresas aseguradoras españolas en el ramo no-vida y, que incorpora como factores de dicho análisis ratios financieros elaborados a partir de los Balances y Cuentas de Resultados de dichas entidades.*

*Para abordar este cometido hemos estructurado nuestro trabajo en las siguientes secciones. En el **capítulo primero** explicamos los principales conceptos del método Rough Set; el **capítulo segundo** ilustra las principales aplicaciones de esta metodología en campos distintos del económico, y resume, los cuatro trabajos más importantes publicados, hasta la fecha, en relación a la predicción de las crisis empresariales; en el **capítulo tercero** describimos el software utilizado para el desarrollo empírico del trabajo, ROSE-Rough Set Data Explorer; el **capítulo cuarto** justifica las variables que vamos a utilizar, los ratios financieros, para pronosticar el fracaso empresarial; en el **capítulo quinto** explicamos los*

*resultados que hemos obtenido y su comparación con una técnica estadística muy utilizada en los problemas de predicción de las crisis empresariales, como es el análisis discriminante. Finalmente, terminamos con las **conclusiones**.*

## **Capítulo 1**

### **EL MÉTODO ROUGH SET**

## I.- INTRODUCCIÓN: ANÁLISIS DE LA DECISIÓN, MODELO DE PREFERENCIAS Y ENFOQUE ROUGH SET

La finalidad de esta introducción no es otra que la de situar el enfoque *rough set* como técnica englobada en el campo de la Inteligencia Artificial y familiarizarnos con la terminología utilizada y su significado dentro de este contexto.

Tomar decisiones es uno de los actos más naturales de los seres humanos. El análisis del proceso de la toma de decisiones ha atraído a los científicos durante mucho tiempo y estos han intentado elaborar un enfoque sistemático y racional para modelizar y resolver los complejos problemas de la toma de decisiones ofreciendo, además, herramientas para abordar dicho proceso.

Mediante el análisis científico de la decisión se pretende sacar a la luz los elementos que intervienen en el proceso de decisión que no son evidentes para los agentes implicados (decisores, accionistas, inversores, expertos, etc.) cuya actitud puede influir en dicha situación.

Precisando aún más, los elementos revelados por el análisis científico de la decisión o explican la situación, o recomiendan, o simplemente favorecen, algún comportamiento para aumentar la coherencia entre las posibilidades ofrecidas por la situación por un lado, y los objetivos y escalas de valores de los agentes particulares implicados (decisores), por otro (Roy, 1985).

En un proceso de toma de decisiones interviene un *decisor* que opera en un *ambiente* dado y toma una *alternativa* de acuerdo a un *sistema de preferencias* (el cual establece las consecuencias asignables a las alternativas, y en particular, a la elegida de entre todas las posibles).

De estos elementos es importante matizar que en este trabajo el decisor conoce el estado de la naturaleza que se le va a presentar, por lo que la alternativa más adecuada será la que le reporte mejores

resultados. En este sentido podríamos tomar los resultados como únicos (porque escogemos los mejores). Todo esto implica que el decisor se encuentra en un ambiente de *certeza* (en caso contrario, el ambiente se denomina de riesgo o incertidumbre).

En términos generales, un problema de decisión implica un conjunto de *objetos* (acciones, estados, competidores, etc.) descritos o valorados por un conjunto de *atributos* (criterios, características, síntomas, etc.). Se puede representar mediante una tabla cuyas filas corresponden a objetos, las columnas a atributos y cada par objeto-atributo se denomina *descriptor*. De forma habitual, uno o varios agentes (expertos, decisores, accionistas, etc.) están implicados en el problema de decisión. Por agente entendemos una persona o una cosa que trabaja para producir un resultado (opinión, decisión, evaluación, etc.). El agente puede estar tanto identificado con un objeto o con un atributo como permanecer “alejado” de la descripción del proceso de decisión. En el primer caso, cada resultado producido por el agente es un descriptor en la tabla mientras que, en el segundo caso, la tabla en su totalidad puede ser formulada por el agente aunque el mismo no esté representado en ella. Pero en ambos casos tanto el objeto, el atributo y los descriptores son conocidos y son únicos dado que nos movemos en un ambiente de certeza.

Los atributos utilizados para describir a los objetos se construyen sobre algunas características elementales de dichos objetos. Los atributos pueden ser *nominales* (también denominados *categoricos* o *cualitativos*) o *cardinales* (también denominados *no-nominales* o *cuantitativos*). Aunque la descripción de un objeto según las características que lo componen puede seguir una ordenación (en el espacio, en el tiempo, o una distribución de probabilidad), normalmente se traduce a un único término, bien sea cuantitativo o cualitativo. Otra posible cualidad de un atributo es el *orden preferencial* de su dominio. Los atributos (finitos o no) ordenados de acuerdo a las preferencias de un agente se convierten en *criterios* permitiendo la comparación de objetos desde puntos de vista particulares.

Si un agente da una descripción de objetos mediante un subconjunto de atributos, dichos atributos se denominan *decisiones*, y los restantes, se denominan *condiciones*, y la tabla en su conjunto se denomina *tabla de decisión*.

El análisis de la decisión intenta contestar a dos cuestiones generales relativas a un problema de decisión. La primera es acerca de la explicación de la propia situación de decisión. Por explicación queremos significar descubrimiento de los hechos y dependencias importantes en la tabla que describe dicha situación de decisión. La segunda cuestión trata de la recomendación de algunas decisiones basándose en el análisis de la información contenida en la tabla.

Es decir, la teoría de la decisión puede orientarse en dos direcciones distintas. Una es la orientación que suele denominarse *positiva (empírica)* que consiste en la elaboración de unos constructos teóricos y articulaciones lógicas que pretenden explicar y predecir el comportamiento de los agentes decisores reales. Otra orientación que suele denominarse *normativa* comienza por definir la racionalidad de los agentes económicos en base a una serie de supuestos justificables intuitivamente. Seguidamente se realizan una serie de operaciones lógicas para deducir el comportamiento óptimo de los agentes decisores como aquel que es compatible con la racionalidad previamente establecida. El enfoque positivo corresponde a una filosofía de *cómo son* (o cómo se comportan) mientras que el enfoque normativo corresponde a una filosofía de *cómo deben de ser* (cómo deben de comportarse) los centros decisores.

Tal y como comprobaremos más adelante, el enfoque rough set es un enfoque fundamentalmente descriptivo de la política seguida en un proceso de toma de decisiones por un decisor o varios. Las reglas de decisión obtenidas al final de la aplicación de esta metodología no van encaminadas a señalar cuál es la decisión óptima, sino la descripción de

la filosofía seguida en dicho proceso de decisión o la explicación (en el sentido mencionado anteriormente) del mismo.

En la toma de decisiones reales, el decisor normalmente toma en consideración muchos puntos de vista (criterios) para la evaluación de las alternativas de decisión. Sin embargo, el problema de toma de decisiones con criterios múltiples no tiene solución sin información adicional sobre las preferencias del decisor. Teniendo la información de las preferencias, uno puede construir sobre la base de criterios múltiples un modelo de preferencia global que contenga “la mejor” solución para un problema de decisión dado.

Un caso importante es aquel en el que la tabla de decisión incluye información sobre las preferencias de un agente. En este caso, el análisis de la información tiende a sintetizar dichas preferencias en un modelo comprensivo que represente la política de toma de decisiones de dicho agente y que pueda ser utilizado para apoyar nuevas decisiones. Dicho modelo es un *modelo global sobre las preferencias*.

Hay dos principales caminos para construir un *modelo global sobre las preferencias* basadas en la información obtenida del decisor. El primero viene del análisis matemático de la decisión y consiste en construir un *modelo funcional o relacional* (Roubens y Vincke, 1985). El *modelo funcional* ha sido muy utilizado dentro del marco de la *teoría de la utilidad multiatributo*. El *modelo relacional* tiene en la forma de *relación de sobreclasificación* ("outranking") o en la *relación fuzzy* sus representaciones más conocidas.

El segundo proviene de la inteligencia artificial y construye el modelo vía *aprendizaje de los ejemplos o adquisición del conocimiento inductiva* (también denominado *inducción de reglas, aprendizaje inductivo*) (Michalski, 1983). El modelo resultante es un conjunto de reglas de la forma Si/Entonces o un árbol de decisión. Este otro camino viene motivado por la hipótesis de que un modelo comprensivo sobre las

preferencias se puede obtener mediante el estudio de las valoraciones racionales efectuadas por los agentes cuando se les presenta un conjunto de objetos representativos de un problema dentro de un dominio de su interés (ejemplos). Lo que subyace dentro de este enfoque es que ofrece más confianza la valoración efectivamente realizada por un agente que la explicación que tuviera que dar sobre la misma dicho agente.

El enfoque *rough set* es un representante típico de esta categoría. También las *redes neuronales* se incluirían dentro de esta categoría.

Profundizando más en esta segunda categoría, resaltamos que, la investigación en la *ingeniería del conocimiento (machine learning)*, una de las áreas de la inteligencia artificial, se ha intensificado mucho en las últimas décadas. El desarrollo ha sido tan extenso que han surgido subáreas enteras dentro de esta gran área.

El mayor esfuerzo investigador se ha realizado en el conocimiento basado en *la similitud*, y más específicamente en una de las subáreas, que es el conocimiento a través de los ejemplos. En esta última el conocimiento se basa en establecer similitudes entre ejemplos positivos, que representan el mismo concepto, y comprobar las no similitudes entre los ejemplos positivos y negativos, siendo estos últimos los que representan otros conceptos. El conocimiento basado en la similitud también se denomina *conocimiento empírico* para subrayar el hecho de que se basa en la inducción del conocimiento que subyace a unos datos empíricos. De este modo, para el método de ingeniería del conocimiento basado en la similitud no es necesario ningún conocimiento a priori, aunque esto no significa que sea totalmente imparcial debido al correspondiente y específico método de aprendizaje (por ejemplo, la forma en que el conocimiento es finalmente expresado).

Con frecuencia, el objetivo de la ingeniería del conocimiento basado en los ejemplos consiste en encontrar una descripción discriminante del concepto. De este modo la tarea consiste en incluir en la



descripción del concepto todos los ejemplos positivos del concepto y excluir de la descripción el conjunto complementario, el cual contendrá todos los ejemplos negativos. Algunos de los sistemas basados en la ingeniería del conocimiento a través de los ejemplos inducen una descripción discriminante mínima.

Ya hemos mencionado que este segundo camino de construir un modelo de preferencia global representa el conocimiento en forma de reglas, árboles de decisión, inducido a través de ejemplos de prueba, y estas son fáciles de comprender por los usuarios finales.

Además, tales reglas pueden ser utilizadas en la práctica mediante su inclusión en una base de conocimiento dentro de un sistema experto que utilice como fuente reglas. Sin embargo después de aprender de los datos de prueba, el conocimiento ha de ser validado. Un ejemplo que no haya sido utilizado en los datos introducidos como prueba puede ser clasificado de manera incorrecta por las reglas. Por tanto, el conocimiento debería ser modificado.

En el contexto de la inteligencia artificial, la información de las preferencias se denomina *conocimiento acerca de las preferencias* y al decisor a menudo se le denomina *experto*.

Por tanto, y retomando los dos caminos vistos anteriormente, en este trabajo nos interesa el segundo modo de construir un modelo de preferencia global, y de entre todos los problemas de decisión nos ocuparemos del problema de *clasificación con criterios múltiples* (aunque también veremos otros problemas de decisión).

Es decir, vamos a tratar *el problema de la clasificación aprendida de la experiencia*. Entendemos experiencia, en un sentido amplio, como un registro de ejemplos, cada cual informando de cómo un objeto dado y descrito por un conjunto de múltiples atributos valorados ha sido clasificado en el pasado.

Asumimos que el número de clases es finito y conocido a priori. Los patrones de clasificación derivados del conjunto de ejemplos constituyen un conocimiento sobre una parte del mundo real o abstracto. Es decir, estamos considerando el caso en el que el experto no expresa directamente el conocimiento sino que éste se deriva de los ejemplos sobre decisiones previas que ha tomado en el pasado sobre un subconjunto de objetos.

El problema de clasificar consiste en asignar cada objeto de un conjunto (descrito por un conjunto de atributos) a una categoría predefinida apropiada (por ejemplo aceptar, rechazar o requerir información adicional). La asignación se realiza sobre la base del conocimiento de un experto. En este caso, el modelo de preferencia global consiste en un conjunto de sentencias lógicas (*reglas de clasificación*) del tipo "si....entonces..." que describen el comportamiento de las preferencias del decisor y relacionan descripciones de los objetos con decisiones sobre su clasificación. El conjunto de reglas se deriva de los ejemplos sobre las decisiones de clasificación tomadas por el decisor (experto) sobre un subconjunto de objetos. El problema de encontrar una representación del conocimiento en forma de reglas de decisión es una parte de la clasificación aprendida de la experiencia. La segunda parte, que es de primordial interés, consiste en utilizar las reglas de decisión para ayudar en las decisiones relativas a la clasificación de *nuevos* objetos.

La clasificación de un nuevo objeto descrito en base a un conjunto de criterios múltiples se sustenta mediante el emparejamiento de su descripción a una de las reglas de clasificación. Si el nuevo objeto encaja exactamente al menos con alguna parte de la condición de una regla de decisión, entonces esta o estas reglas se presentan al decisor junto con la información sobre el número de ejemplos que apoyan cada posible regla. En otros casos, se presentan al decisor el conjunto de reglas "más cercanas" a la descripción del nuevo objeto. La noción de "cercanía"

implica la utilización de una medida de distancia. En este trabajo describiremos las dos principales medidas de distancia (Métrica  $L_p$  y la Relación de proximidad valorada) que utiliza esta metodología y que se incorporan a los softwares que desarrollan la misma.

Hasta ahora al hablar de las preferencias del decisor no hemos tenido en cuenta que la misma suele ser imprecisa (inconsistente) debido a diferentes fuentes de imprecisión, como son las inconsistencias de los ejemplos. Esta inconsistencia en la descripción de los objetos lleva a la ambigüedad en su clasificación, es decir objetos descritos por los mismos valores de los atributos (de condición) pero que sin embargo están asignados a diferentes clases (atributos de decisión) ( Roy, 1989). Esta incertidumbre impide una clasificación precisa de los objetos y la búsqueda de dependencias entre valores de los atributos de condición y atributos de decisión.

En consecuencia, el sistema de conocimiento es a menudo obligado a tratar con imprecisiones o trabaja en presencia de "ruido". Hay dos razones principales que justifican su existencia: que la evidencia sea incompleta o sea contradictoria. Desde el punto de vista de nuestro trabajo nos interesa la inducción de reglas en el caso de que la evidencia sea contradictoria, esto es, que los ejemplos o casos que tenemos sean inconsistentes. En la práctica, eso significa que dados dos ejemplos, clasificados por los mismos valores de los atributos, cada uno de ellos pertenece a conceptos o clases diferentes.

En realidad, esa imprecisión proviene de la "granularidad" existente en un sistema de información que represente las preferencias de un decisor dado. Debido a esta granularidad, las reglas que describen un comportamiento que emana de las preferencias del decisor, pueden clasificarse en *determinísticas* y *no-determinísticas*. Las reglas son determinísticas si pueden describir de forma unívoca (por medio de "gránulos" del sistema de información que representa las preferencias) un comportamiento, y son no determinísticas en cualquier otro caso.

Muchos de los sistemas de conocimiento existentes que manejan la incertidumbre, se basan en la teoría de la probabilidad (algunos utilizan el enfoque Bayesiano y otros utilizan herramientas estadísticas). Sin embargo uno de los métodos, y en el que no vamos a entrar, más populares es inducir árboles de decisión.

Pawlak (1982), introdujo a comienzos de los años ochenta una herramienta formal, matemática, para tratar con la incertidumbre o la vaguedad inherente a un proceso de decisión, es decir, dio un marco formal en el cual las inconsistencias que aparecen en una tabla de decisión no se corrigen sino que en su lugar se obtienen, según veremos, reglas determinísticas y no determinísticas para una representación dada del conocimiento que es la denominada teoría del *rough set*.

Esta teoría implica el cálculo de particiones, divisiones o clases, según queramos. Es algo diferente tanto de la teoría estadística de la probabilidad como de la teoría *fuzzy set*. Para comprender en qué se diferencian describiremos las tres categorías generales de imprecisión que encontramos en el análisis científico. La primera categoría ocurre, cuando un acontecimiento es aleatorio en naturaleza (por ejemplo, lanzar una moneda) y por lo tanto no se sabe qué resultado en particular (de todos los posibles) es el que se va a materializar. Por lo tanto, la imprecisión asociada con este tipo de acontecimientos puede describirse por la teoría estadística de la probabilidad. La segunda surge del hecho de que los objetos pueden no pertenecer exclusivamente a una única categoría pero pueden pertenecer a varias categorías aunque con diferentes grados, en este caso la imprecisión toma forma de pertenencia difusa a un conjunto (el ejemplo más claro es la clasificación de los objetos de color gris en las categorías “negro” o “blanco”). Esta segunda categoría es el objeto de la lógica *fuzzy*. La tercera categoría es la teoría *rough set* que es útil cuando las clases en las que han de clasificarse los objetos son imprecisas, pero sin embargo pueden aproximarse mediante conjuntos precisos (Nurmi et al., 1996).

Estas diferencias muestran una de las principales ventajas de la teoría rough set y es que no necesita ninguna información adicional acerca de los datos como puede ser una distribución de probabilidad en estadística o el grado o probabilidad de pertenencia en la teoría *fuzzy set*.

Esta teoría es una herramienta para trabajar en condiciones en las cuales nuestro conocimiento es inexacto, incierto o vago para aplicaciones de la inteligencia artificial, como por ejemplo ocurre en el campo de la medicina o para conocer patrones de conocimiento.

El enfoque rough set presenta las siguientes ventajas:

- analiza los hechos escondidos en los datos,
- no necesita ninguna información adicional o preliminar sobre los datos como probabilidades a priori en la teoría de la probabilidad o grado de pertenencia en la teoría fuzzy set,
- reduce los atributos y objetos superfluos para obtener los denominados *reductos* (o *conjuntos mínimos de atributos*) que aseguran la misma calidad de clasificación que el conjunto total.
- obtiene la mínima representación del conocimiento en forma de reglas de decisión teniendo en consecuencia, una interpretación sencilla,
- está especialmente indicada para tratar las inconsistencias de los datos; no corrige las inconsistencias que se presentan en los ejemplos o en los casos sino que calcula aproximaciones "por arriba" y "por abajo". En consecuencia, y basándose en estas aproximaciones, se inducen dos conjuntos de reglas: determinísticas (ciertas) y no-determinísticas (posibles),
- conceptualmente es simple y por tanto requiere algoritmos simples.

Las ideas que vamos a exponer en las secciones siguientes de este capítulo son algunos conceptos básicos de la teoría Rough Set descritos por Pawlak 1982, 1984 y 1991 y seguiremos su exposición. Aunque en la actualidad existen extensiones de esta teoría (Greco et al., 1998), expondremos el enfoque clásico, que es el que hemos seguido en nuestra aplicación práctica.

Básicamente, el *nuevo enfoque*, incorpora información adicional acerca de las propiedades de ordenación de los atributos analizados, en consecuencia se obtiene potencialmente como resultado un modelo final más compacto ya que algunas reglas contradictorias para determinados casos son de este modo eliminadas (McKee, 2000). En otras palabras, utiliza información adicional para generar un modelo final más simple.

El enfoque original no ordena el dominio de los atributos. Asume que los diferentes valores de los mismos atributos son igualmente preferibles y que únicamente el valor predictivo del atributo, tal y como es revelado por los datos, será tenido en cuenta en el modelo. Por lo tanto, esto le hace menos restrictivo en cuanto a las hipótesis sobre los datos que el nuevo enfoque.

## II.- APROXIMACIÓN AL CONCEPTO DE ROUGH SET

La teoría rough set (conjunto impreciso/aproximado), tal y como se ha mencionado, fue introducida por Pawlak en 1982.

La hipótesis fundamental en la teoría rough set es la siguiente: los objetos del universo se perciben únicamente a través de la información que hay accesible acerca de ellos, esto es, a través de valores de atributos es como podemos evaluar estos objetos.

La filosofía del método se basa en la suposición de que con cada objeto del universo que estamos considerando se puede asociar alguna información (datos, conocimiento). Los objetos caracterizados por la misma información no son discernibles en vista de la información disponible. La relación de no diferenciación generó de este modo la base matemática para esta teoría.

En consecuencia, la observación de que los objetos pueden no ser discernibles en términos de descriptores (valor de un atributo para un objeto) debido a la información imprecisa es el punto de partida de la filosofía rough set. Esta información imprecisa es la causa de la no diferenciación de los objetos en términos de datos disponibles y evita, en consecuencia, su asignación precisa a un conjunto.

Cualquier conjunto de objetos que no son discernibles se llama *conjunto elemental*, y forma un "gránulo" básico del conocimiento acerca del universo. Cualquier conjunto de objetos que es la unión de algunos conjuntos elementales se considera como "preciso" y en el caso contrario el conjunto se dice "rough" (impreciso, vago). Introducido el concepto "rough" hablaremos de aquí en adelante de rough set. Por tanto, y de manera intuitiva, un rough set es un conjunto de objetos que, en general, no pueden ser caracterizados de manera precisa en términos de valores de un conjunto de atributos.

Por lo tanto, la clasificación de los objetos se basa en la información que hay accesible sobre ellos y no en los objetos en sí mismos. Normalmente junto con la información existente sobre los objetos se da una clasificación suministrada por un experto. El problema de clasificación en este caso consiste en qué medida es posible reflejar mediante atributos (de condición) la clasificación efectuada por el experto.

Pero nos encontramos con un problema para realizar la clasificación, y es que la no diferenciación de objetos por medio de atributos impide generalmente su asignación precisa a un conjunto, tal y como se ha mencionado. De este modo, dada una relación de equivalencia (clasificación) en la que hay una relación de no diferenciación (los objetos no son discernibles), se induce una *aproximación* del espacio de las clases de equivalencia (clasificaciones).

La aproximación del espacio y la aproximación de un conjunto o de una familia de conjuntos (en particular, clasificaciones) en este espacio, son dos conceptos próximos de la teoría rough set de gran importancia práctica: un rough set es un par de aproximaciones por arriba y por debajo de un conjunto en términos de objetos no discernibles. En otras palabras, un rough set es una colección de objetos que, en general, no pueden ser clasificados de manera precisa en términos de los valores del conjunto de atributos, mientras que las aproximaciones por arriba y por abajo sí pueden.

En consecuencia, cada rough set tiene casos fronterizos, esto es objetos que no pueden clasificarse con certeza como miembros del conjunto o de su complementario.

En definitiva, un rough set puede ser reemplazado o representado por un par de conjuntos precisos, llamados *la aproximación por encima* y *por debajo*. La aproximación por debajo consiste en todos los objetos que con seguridad pertenecen al conjunto y la aproximación por encima contiene los objetos que posiblemente pertenecen al conjunto.



Usando una aproximación por encima y por debajo de un conjunto (o familia de conjuntos – partición) se puede definir la *precisión* de la aproximación y la *calidad* de la misma, que son números del intervalo  $[0, 1]$  y que definiremos en las próximas secciones.

Utilizaremos el enfoque del rough set para extraer reglas de clasificación aplicadas al problema del fracaso empresarial; en las secciones próximas caracterizaremos la metodología, incluyendo el uso de reglas de clasificación para apoyar la decisión y posteriormente describiremos los tipos de problemas de decisión a los que se puede aplicar esta metodología.

### III.- TABLA DE INFORMACIÓN Y RELACIÓN DE NO DIFERENCIACIÓN

Una premisa central de la filosofía rough set es que el *conocimiento* consiste en la capacidad de *clasificación*. Un comportamiento racional de cualquier agente en el ámbito externo se basa en la capacidad para clasificar objetos reales o abstractos. Para clasificar, uno debe percibir algunas diferencias entre los objetos, y de esta manera se pueden formar clases de objetos que no presentan diferencias notables. Estas clases de objetos que no se pueden diferenciar pueden considerarse como los *pilares (conceptos)* utilizados para construir un conocimiento sobre el mundo real o abstracto.

De aquí en adelante, el conocimiento se define como una familia de partes de un conjunto fijo (universo)  $U$  o, lo que es lo mismo desde el punto de vista matemático, como una familia de clases de equivalencia  $R$  en  $U$ . Es decir, un *conocimiento base*,  $k = (U, R)$ , se compone de  $U$  y de una familia  $R$  de partes (o relaciones de equivalencia) sobre  $U$ .

Esta particular visión del conocimiento es de naturaleza semántica donde la granularidad del conocimiento (no diferenciación de algunos objetos) es de primera importancia y se usará para definir los conceptos claves de la teoría rough set: *aproximación, dependencia y reductos*.

La teoría del rough set asume la representación del conocimiento de los objetos en forma de una *tabla de información*, que es un caso especial de un sistema de información. En las filas de la tabla se indican los *objetos* (acciones, alternativas, candidatos, pacientes, empresas, etc.), mientras que las columnas se corresponden con los *atributos*. Las entradas en la tabla son los *valores del atributo*. Es decir, la entrada en columna  $q$  y en fila  $x$  tiene el valor  $f(x, q)$ . Cada fila en la tabla representa información sobre un objeto en  $S$ , siendo  $S$  el sistema de información.

El sistema de información también se denomina *sistema de representación del conocimiento*. La teoría rough set utiliza la noción de atributo en lugar de la de criterio porque el primero es más general que el segundo debido a que el dominio (escala) de un criterio ha de ordenarse de menor a mayor preferencia mientras que el dominio de los atributos no ha de ser ordenado; la noción de criterio puede también usarse cuando la ordenación preferencial del dominio de un atributo es importante en un contexto dado. Esto no es sólo válido para los atributos de condición; de igual modo, el dominio de un atributo de decisión puede ser ordenado o no.

Por tanto, para cada par (objeto, atributo) se conoce un valor denominado *descriptor*. Cada fila de la tabla contiene descriptores que representan información correspondiente a un objeto del universo, es decir, cada fila en la tabla representa la información acerca de un objeto en S. Por ejemplo, un fichero de datos relativo a pacientes que sufren de una determinada enfermedad.

Además, y en general, el conjunto de atributos se parte en un subconjunto de *atributos de condición* (características, variables, rasgos,...) y en subconjunto de *atributos de decisión* (opiniones, clases, diagnósticos,...). Las *dependencias causa-efecto* entre estos dos grupos de atributos es lo más interesante del análisis de los sistemas de información. La otra parte esencial se refiere a la *posible reducción* de todos los atributos y casos superfluos en el sistema de información y la representación de las dependencias entre los valores de los atributos en forma de *reglas de decisión*.

Formalizando lo expresado con anterioridad, tendríamos:

Por una *tabla de información* entendemos el conjunto  $S = \langle U, Q, V, f \rangle$  donde:

$U$  es un conjunto finito de objetos (observaciones, estados, casos, acontecimientos,...), el Universo que consideramos.

$Q$  es un conjunto finito de *atributos* (características, variables,...)

$$V = \bigcup_{q \in Q} V_q \quad (V_q \text{ es el dominio del atributo } q) \text{ y}$$

$f: U \times Q \rightarrow V$  es una función total tal que  $f(x, q) \in V_q$  para cada  $q \in Q$  y  $x \in U$ , es llamada *una función de información* (Pawlak, 1991). Cualquier par  $(q, v)$ ,  $q \in Q$  y  $v \in V_q$ , se denomina descriptor en  $S$ .

### Relación de no diferenciación

Sea  $S = \langle U, Q, V, f \rangle$  una tabla de información y sea  $P$  un subconjunto de atributos de  $Q$  ( $P \subseteq Q$ ) y  $x, y \in U$ . Decimos que  $x$  e  $y$  *no son discernibles* para el conjunto de atributos  $P$  en  $S$  si, y solo si,  $f(x, q) = f(y, q)$  para cada  $q \in P$ . Todo  $P \subseteq Q$  genera una relación binaria en  $U$  llamada *relación de no diferenciación de  $P$* , y se denota  $I_P$  que es una relación de equivalencia para cualquier  $P$ . Las clases de equivalencia de la relación  $I_P$  se llaman *conjuntos elementales  $P$*  (conceptos o categorías del conocimiento básico de  $P$ ) en  $S$ . La familia de las clases de equivalencia de la relación  $I_P$  en  $U$  se denomina  $U/I_P$  o, de forma abreviada,  $U/P$  y también se le conoce por *conocimiento básico de  $P$*  en  $S$  y por  $I_P(x)$  se denota el conjunto elemental  $P$  (clase de equivalencia de  $I_P$ ) que contiene al objeto  $x \in U$ .

Los conjuntos elementales  $Q$  en  $S$  se denominan **átomos** y se crean por las clases de equivalencia.

Ya que todos los objetos que están en la misma clase de equivalencia no se pueden diferenciar, deben tener la misma descripción.

Una *descripción* es un conjunto de pares atributo-valor. Algunas veces es más natural considerar una descripción como la conjunción lingüística de pares atributo-valor.

Si se distingue entre *atributos de condición* y *atributos de decisión* en un sistema de información, obtenemos, como veremos una tabla de decisión.

Para clarificar los conceptos que iremos explicando, vamos a ilustrar este capítulo con un ejemplo sobre la selección de alumnos para un colegio (Moscarola, 1978 y Slowinski, 1993).

Los candidatos han remitido sus solicitudes con sus calificaciones académicas, su curriculum vitae y las referencias del colegio anterior. Todos estos documentos son examinados por un comité que decidirá su admisión o no. Basándose en los documentos anteriores, los candidatos son *descritos* utilizando siete criterios (atributos de condición). A continuación presentamos la lista de estos criterios y sus escalas correspondientes, ordenadas de mejor a peor:

$C_1$  – nota de matemáticas, {5, 4, 3}

$C_2$  – nota de física, {5, 4, 3}

$C_3$  – nota de inglés, {5, 4, 3}

$C_4$  – nota media en el resto de asignaturas, {5, 4, 3}

$C_5$  – tipo colegio donde realizó la educación secundaria, {1, 2, 3}

$C_6$  – motivación, {1, 2, 3}

$C_7$  – opinión del colegio anterior, {1, 2, 3}

Quince candidatos son los que han presentado las solicitudes y han sido clasificados por el comité después de la debida consideración. Estos candidatos son los ejemplos que vamos a utilizar para conocer cuales son las *preferencias* de dicho comité a la hora de admitir o rechazar a un candidato.

El atributo de decisión es  $d$  y en base al mismo se puede hacer una clasificación dicotómica de los candidatos:  $d= A$  significa admisión y  $d= R$  significa no admisión.

Al haber distinguido entre atributos de condición y atributos de decisión en este sistema de información, obtenemos, una tabla de decisión.

La tabla de decisión sería la siguiente:

<b>Criterios</b>	<b>C<sub>1</sub></b>	<b>C<sub>2</sub></b>	<b>C<sub>3</sub></b>	<b>C<sub>4</sub></b>	<b>C<sub>5</sub></b>	<b>C<sub>6</sub></b>	<b>C<sub>7</sub></b>	<b>D</b>
<b>Candidatos</b>								
<b>X<sub>1</sub></b>	4	4	4	4	2	2	1	A
<b>X<sub>2</sub></b>	3	3	4	3	2	1	1	R
<b>X<sub>3</sub></b>	3	4	3	3	1	2	2	R
<b>X<sub>4</sub></b>	5	3	5	4	2	1	2	A
<b>X<sub>5</sub></b>	4	4	5	4	2	2	1	A
<b>X<sub>6</sub></b>	3	4	3	3	2	1	3	R
<b>X<sub>7</sub></b>	4	4	5	4	2	2	2	A
<b>X<sub>8</sub></b>	4	4	4	4	2	2	2	A
<b>X<sub>9</sub></b>	4	4	4	4	2	2	2	R
<b>X<sub>10</sub></b>	5	3	5	4	2	1	2	A
<b>X<sub>11</sub></b>	5	4	4	4	1	1	2	A
<b>X<sub>12</sub></b>	5	3	4	4	2	2	2	A
<b>X<sub>13</sub></b>	4	3	3	3	3	2	2	R
<b>X<sub>14</sub></b>	3	3	4	3	2	3	3	R
<b>X<sub>15</sub></b>	4	5	5	4	2	1	1	A

Esta claro que el conjunto de los atributos de condición es:

$$C = \{ C_1, C_2, C_3, C_4, C_5, C_6, C_7 \}$$

y el conjunto de los atributos de decisión es:

$$D = \{d\}$$

Denominemos como  $Y_A$  al conjunto de candidatos admitidos y a  $Y_N$  al conjunto de los rechazados por el comité. Por tanto,  $Y_A = \{x_1, x_4, x_5, x_7, x_{10}, x_{11}, x_{12}, x_{15}\}$ ,  $Y_N = \{x_2, x_3, x_6, x_8, x_9, x_{13}, x_{14}\}$ ,  $Y = \{Y_A, Y_N\}$ . Los conjuntos  $Y_A$  e  $Y_N$  son clases definidas por el conjunto  $D$  en la tabla de decisión.

#### IV.- APROXIMACIÓN DE CONJUNTOS

Ya hemos mencionado como un conjunto finito de objetos (universo)  $U$  y una familia finita de relaciones de equivalencia  $R$  sobre  $U$  (es decir, el conocimiento acerca de  $U$ ) puede expresar formalmente como el conocimiento  $R$  acerca de  $U$  a través de un sistema relacional  $K=\langle U, R \rangle$ . El principal problema en el que estamos interesados es el siguiente: Dado un subconjunto  $X \subseteq U$  y el conocimiento  $K=\langle U, R \rangle$ , expresar las propiedades de  $X$  en términos del conocimiento disponible.

Debido a que la intersección teórica de las relaciones de equivalencia es también una relación de equivalencia, la familia resultante de clases de equivalencia (partición) puede verse como una familia de conjuntos elementales (átomos, gránulos) de conocimiento  $K=\langle U, R \rangle$ . Por tanto, a partir de ahora nuestro problema se reduce a definir el subconjunto  $X$  en términos de conjuntos elementales, esto es, en representarlo como unión de átomos. Sin embargo, esto no es siempre posible, por tanto hay que introducir dos conceptos, *la aproximación por arriba* y *por abajo*.

La aproximación por debajo de  $X$  en  $K=\langle U, R \rangle$  es la unión de los conjuntos elementales que están incluidos en  $X$ , (es decir, la aproximación por debajo indica cuales de las clases de equivalencia de la relación  $R$  son subconjuntos de  $X$ , o sea, todos sus elementos son también elementos de  $X$ ). La aproximación por arriba de  $X$  en  $K=\langle U, R \rangle$  es la unión de todos los conjuntos elementales cuya intersección con  $X$  no da el conjunto vacío (por tanto, indica aquellas clases de  $K$  que tienen al menos un elemento común con  $X$ ). En otras palabras, la aproximación por debajo de  $X$  se compone de aquellos subconjuntos que son necesariamente también subconjuntos de  $X$ , mientras que la aproximación por arriba se compone de aquellos subconjuntos que contienen algunos elementos de  $X$ .



Estas aproximaciones corresponden, respectivamente, con el máximo número de objetos incluidos que *con seguridad* pertenezcan a X y con el conjunto de objetos que *posiblemente* pertenezcan a X. La diferencia entre la aproximación por arriba y por debajo es el *conjunto frontera* que está formado por todos los objetos que no pueden clasificarse con certeza en X o en su complementario. El número cardinal del conjunto frontera nos indica con qué exactitud se puede describir X en términos del conocimiento disponible.

Acabamos de ver de una manera intuitiva en qué va a consistir el problema de la aproximación de conjuntos. Ahora vamos a formalizar dicha cuestión.

Dado un sistema de información S y dados  $P \subseteq Q$  e  $Y \subseteq U$ . Con cada subconjunto y con una relación de equivalencia asociamos dos subconjuntos, las denominadas aproximaciones por encima y por debajo.

Nuestra finalidad no es otra que verificar en qué medida el conjunto P describe los objetos de un determinado conjunto Y.

La aproximación por debajo de Y, denotada por  $\underline{P}Y$ , y la aproximación por encima de Y, denotada por  $\overline{P}Y$ , se definen como:

$$\underline{P}Y = \{x \in U: I_p(x) \subseteq Y\},$$

$$\overline{P}Y = \{x \in U: I_p(x) \cap Y \neq \emptyset\} = \bigcup_{x \in Y} I_p(x)$$

La frontera P (también denominada *región de duda*) del conjunto Y, y denotada por  $B_{n_p}$ , se define como:

$$B_{n_p}(Y) = \underline{P}Y - \overline{P}Y \text{ (donde "-" es el operador diferencia de conjuntos)}$$

El conjunto  $\underline{P}Y$  es el conjunto de todos los elementos de U que pueden ciertamente clasificarse como elementos de Y, utilizando el

conjunto de atributos P. El conjunto  $\overline{PY}$  es el conjunto de los elementos de U que pueden posiblemente clasificarse como elementos de Y, utilizando el conjunto de atributos P. El conjunto  $Bn_p(Y)$  es el conjunto de elementos que no pueden ser con certeza clasificados en Y utilizando el conjunto de atributos P.

Con cada conjunto  $Y \subseteq U$ , se asocia una *precisión de la aproximación* del conjunto Y por P en S, o en pocas palabras, la precisión de Y, se define como:

$$\alpha_p(U) = \frac{\text{card}(\underline{PY})}{\text{card}(\overline{PY})}$$

donde card () es el número cardinal de un conjunto (Slowinski y Stefanowski, 1994a, 3-4).

Por tanto, aquellos subconjuntos de un sistema de información que no pueden ser expresados exactamente mediante el empleo de atributos disponibles, pueden definirse de manera imprecisa utilizando las aproximaciones por encima y por debajo. Sin embargo, también pueden definirse de manera precisa si  $\underline{PY} = \overline{PY}$ .

Continuando con el ejemplo iniciado en las secciones anteriores en nuestra tabla existen 13 conjuntos elementales que son: dos pares de candidatos que no se diferencian,  $\{x_4, x_{10}\}$ ,  $\{x_8, x_9\}$  y 11 candidatos que sí se diferencian. Las aproximaciones por arriba y por debajo de los conjuntos  $Y_A$  e  $Y_N$  son: respectivamente:

$$\underline{CY}_A = \{x_1, x_4, x_5, x_7, x_{10}, x_{11}, x_{12}, x_{15}\}$$

$$\overline{CY}_A = \{x_1, x_4, x_5, x_7, x_8, x_9, x_{10}, x_{11}, x_{12}, x_{15}\}$$

$$Bn_p(Y_A) = \{x_8, x_9\}$$

$$\underline{CY}_R = \{x_2, x_3, x_6, x_{13}, x_{14}\},$$

$$\overline{CY}_R = \{x_2, x_3, x_6, x_8, x_9, x_{13}, x_{14}\}$$

$$B_{\eta_p}(Y_R) = \{x_8, x_9\}$$

La precisión de la aproximación para cada conjunto es la siguiente:

$$\alpha(Y_A) = 8/10=0.8 \text{ y } \alpha(Y_N) = 5/7 =0.71.$$

Observemos que la región de duda o frontera está compuesta por dos candidatos, el 8 y el 9. Ambos tienen la misma puntuación de acuerdo a los criterios definidos por el conjunto C pero el comité ha admitido al 8 y rechazado al 9. Esto implica que la decisión no es consistente con la evaluación de los candidatos en base a los criterios considerados. Por tanto, en apariencia, el comité debería recabar información adicional de las solicitudes de los candidatos o entrevistarlos. Esta conclusión lleva consigo que el comité o bien adopta un criterio adicional discriminatorio o bien, si su definición es demasiado difícil, debería crear una tercera categoría de candidatos: aquellos a los que se debe entrevistar.

## V.- APROXIMACIÓN DE LA PARTICIÓN DE $U$ O CLASIFICACIÓN ROUGH

Sea  $S$  una tabla de información,  $P \subseteq Q$  (es decir, un subconjunto de atributos de  $Q$ ) y sea  $Y = \{Y_1, Y_2, \dots, Y_n\}$  una clasificación (o partición) de  $U$  en  $S$ , esto es,  $Y_i \cap Y_j = \emptyset, \forall i, j \leq n, i \neq j$  y  $\bigcup_{i=1}^n Y_i = U$ .

El origen de esta clasificación es independiente de los atributos de  $P$ . Se obtiene, por ejemplo, al resolver un problema de clasificación efectuado por un experto.

Por aproximaciones por debajo y por encima de  $Y$  en  $S$  se denomina el conjunto  $\underline{P}Y = \{\underline{P}Y_1, \underline{P}Y_2, \dots, \underline{P}Y_n\}$  y  $\overline{P}Y = \{\overline{P}Y_1, \overline{P}Y_2, \dots, \overline{P}Y_n\}$  respectivamente. El coeficiente :

$$\gamma_P(Y) = \frac{\sum_{i=1}^n \text{card}(\underline{P}Y_i)}{\text{card}(U)}$$

es llamado la *calidad de la aproximación de la clasificación o partición*  $Y$  por el conjunto de atributos  $P$ , o en otras palabras, *calidad de clasificación*. Expresa el ratio de todos los objetos  $P$  correctamente clasificados respecto a todos los objetos del sistema (Slowinski y Stefanowski, 1994a, 6-8).

En nuestro ejemplo la calidad tomaría el siguiente valor:

$$\gamma_P(Y) = (8+5)/15 = 0.87.$$

Con los conceptos de precisión y calidad de la aproximación vistos en las dos últimas secciones podemos definir en qué medida está descrito el conjunto (o clasificación) de objetos examinado utilizando la información disponible. En consecuencia, podemos evaluar la importancia de cada atributo en particular para la clasificación ya que si la calidad

toma un valor bajo significa que hay que tomar en consideración otros atributos no considerados a la hora de poder tomar una decisión.

## VI.- REDUCCIÓN Y DEPENDENCIA DE ATRIBUTOS

Una de las funciones más importantes de la investigación rough set es el descubrimiento de las dependencias entre atributos al analizar una tabla de información. Mediante dicho descubrimiento se pretende lograr la reducción del conocimiento, es decir, el proceso de expresar un conocimiento dado del modo más eficiente.

Descubrir las *dependencias* en  $K=\langle U, R \rangle$  consiste en encontrar las relaciones entre las particiones, es decir, las relaciones de equivalencia pertenecientes a la familia  $R$ . En otras palabras, la dependencia nos indica cómo algunos conceptos del conocimiento  $K=\langle U, R \rangle$  pueden expresarse mediante otros conceptos del conocimiento  $K=\langle U, R \rangle$ .

Observemos que para el conocimiento  $K=\langle U, R \rangle$ , dos familias diferentes de relaciones de equivalencia  $R$  y  $R'$  (siendo  $R'$  una subfamilia de  $R$ ), pueden generar la misma familia de conjuntos elementales. Entonces, es importante conocer si podemos *reducir* aún más  $R'$  manteniendo la familia de conjuntos elementales, es decir, sin perder una parte del conocimiento. Introducimos, de esta manera, el concepto de *reducto*: es el mínimo subconjunto de atributos que mantienen la misma familia de conjuntos elementales que el conjunto de todos los atributos.

Formalizando lo expresado en los párrafos precedentes, tenemos que el conjunto de atributos  $R \subseteq Q$ , depende del conjunto de atributos  $P \subseteq Q$  en  $S$  (denotado por  $P \rightarrow R$ ) si y solo si  $I_p \subseteq I_r$ . De manera intuitiva,  $R$  depende de  $P$  si los valores de los atributos de  $R$  están determinados únicamente por los valores de los atributos de  $P$ , es decir, existe una dependencia funcional entre los valores de  $R$  y de  $P$ .

Descubrir las dependencias entre los atributos permite la reducción del conjunto de atributos. El subconjunto  $P \subseteq Q$  es *independiente* en  $S$  si, y solo si, para cada  $P' \subset P$ ,  $I_{P'} \not\subseteq I_P$ ; en cualquier otro caso  $P \subseteq Q$  es *dependiente* en  $S$ .

En las aplicaciones prácticas estamos interesados en reducir aquellos atributos que son redundantes en  $S$  (es decir, estamos interesados en obtener lo que hemos denominado *reductos*). El subconjunto  $P \subseteq Q$  es un reducto de  $Q$  en  $S$  si  $P$  es el mayor de los conjuntos independientes en  $Q$ . Como veremos ahora, para encontrar un reducto se puede utilizar el concepto de *calidad de la aproximación*  $\gamma_p(Y)$ , donde  $Y$  es una clasificación o partición de  $U$ , tal y como hemos indicado en la sección anterior.

La reducción de atributos se consigue mediante la investigación de un modelo tal que el conjunto reducido de atributos proporcione la misma calidad de clasificación que el conjunto original de atributos  $Q$ . El mínimo subconjunto de atributos  $R \subseteq P \subseteq Q$  tal que  $\gamma_P(Y) = \gamma_R(Y)$  (es decir, que asegure la misma calidad de clasificación que el conjunto de todos los atributos) es denominado *reducto- $Y$*  de  $P$  (o, simplemente, reducto si no hay ambigüedad al sobrentender de  $Y$ ) y es denotado por  $RED_Y(P)$ . También se le denomina *conjunto mínimo* (o subconjunto) en  $S$  o simplemente *conjunto mínimo de atributos*.

Vamos a dar una propiedad que establece la relación entre los reductos y la dependencia:

Si  $R'$  es un reducto de  $R$ , entonces  $R' \rightarrow R - R'$ . Este tipo de relación se denomina *dependencia básica* entre atributos. Es obvio que, si  $P \rightarrow R$  entonces  $P \rightarrow q$  para cada  $q \in R$ . Esta clase de dependencia se denomina *elemental*.

En el caso en que la tabla de información tenga más de un reducto de  $Y$ , la intersección de todos ellos se llama *Núcleo- $Y$* , que denotaremos por  $Core_Y(P)$

Su valor es:

$$Core_Y(P) = \bigcap RED_Y(P).$$

El núcleo es una colección de los atributos más relevantes en la tabla que no pueden ser eliminados de S sin que disminuya la calidad de la aproximación de la clasificación Y.

En otras palabras, dentro de la reducción del conocimiento, los conceptos de núcleo y reducto pueden tomar la siguiente interpretación. El núcleo se compone de aquellas clasificaciones que son las más esenciales en el conocimiento. Ninguna relación del núcleo puede ser eliminada en la reducción del conocimiento sin distorsionar el mismo. Un reducto, por otro lado, proporciona un conjunto de relaciones que es suficiente para caracterizar el conocimiento sin perder nada esencial. Es decir, los reductos son intercambiables (se puede considerar uno u otro) pero el núcleo no.

Sin entrar en la casuística de modelos propuestos para hallar los reductos y el núcleo, presentaremos brevemente uno de los más utilizados.

Skowron propuso representar el conocimiento en la forma de una *matriz de diferenciación* (Skowron, 1991a y 1991b). Esta representación tiene muchas ventajas, en particular permite de manera simple el cálculo del núcleo, reductos y otros conceptos considerados.

La idea es la siguiente:

Sea  $S = (U, Q, V, f)$  un sistema de representación del conocimiento (tabla de información) con  $U = \{x_1, x_2, \dots, x_n\}$ . Por matriz de diferenciación de S, que denominaremos  $M(S)$ , tenemos la matriz  $n \times n$  definida del siguiente modo:

$$c_{ij} = \{q \in Q: f(x_i, q) \neq f(x_j, q)\} \text{ para } i, j = 1, 2, \dots, n.$$

La entrada  $c_{ij}$  es el conjunto de todos los atributos que diferencian todos los objetos  $x_i$  y  $x_j$ .



El núcleo puede definirse ahora como el conjunto de todas las entradas de un único elemento de la matriz de diferenciación, esto es:

$$\text{CORE}(U) = \{q \in Q: c_{ij} = (q), \text{ para algún } i, j\}.$$

Puede fácilmente comprobarse que  $R \subseteq Q$  es el reducto de  $Q$ , si  $R$  es el mínimo subconjunto de  $Q$  tal que  $R \cap c \neq \emptyset$  para cualquier entrada no vacía  $c$  ( $c \neq \emptyset$ ) en  $M(S)$ .

En otras palabras el reducto es el mínimo conjunto de atributos que diferencian todos los objetos a distinguir por la totalidad de todos los atributos.

La matriz de diferenciación es simétrica lo cual simplifica y facilita su cálculo.

De igual modo se puede calcular el núcleo relativo y los reductos, teniendo ahora únicamente en mente que lo que hay que diferenciar son clases de equivalencia generadas por atributos de decisión en lugar de como hasta ahora, objetos, mediante el empleo de atributos de condición. En esto se basan, como veremos, algunos de los algoritmos utilizados para extraer las reglas de decisión.

Para aclarar el concepto de matriz de diferenciación proponemos el siguiente ejemplo (Pawlak,1982). Consideremos que nuestra tabla de información es la siguiente:

U	a	b	c	d
1	0	1	2	0
2	1	2	0	2
3	1	0	1	0
4	2	1	0	1
5	1	1	0	2

Donde a, b, c, y d son los criterios y 1,2,3,4,5 son los candidatos evaluados por los criterios.

La matriz sería la siguiente:

	1	2	3	4	5
1					
2	a, b, c, d				
3	a, b, c,	b, c, d			
4	a, c, d	a, b, d	a, b, c, d		
5	a, c, d	b	b, c, d	a, d	

De la tabla se deduce fácilmente que el núcleo es el atributo  $b$  y que en la tabla hay dos reductos el  $\{a, b\}$  y el  $\{d, b\}$ .

En nuestro ejemplo de los candidatos  $y$ , procediendo de la misma manera, obtendríamos los siguientes reductos:

$$RED_{y1}(C) = \{c_2, c_3, c_6, c_7\}$$

$$RED_{y2}(C) = \{c_1, c_3, c_7\}$$

$$RED_{y3}(C) = \{c_2, c_3, c_5, c_7\}$$

Por tanto, podría decirse que el comité toma las decisiones teniendo en cuenta únicamente los criterios de los reductos y descartando el resto. Por ejemplo, el criterio  $c_4$  no influye en ninguna decisión puesto que no aparece en ningún reducto.

El núcleo está formado por la intersección de los tres reductos, es decir:

$$CORE_y(C) = \{c_3, c_7\}$$

## VII.- REGLAS DE DECISIÓN

En esta sección vamos a tratar la inducción de reglas de decisión de un conjunto de datos que representan la experiencia en un dominio. Se supone que el conjunto de datos contiene información de un conjunto de *objetos* descritos por un conjunto de *atributos*. El problema, por tanto, consiste en encontrar reglas que determinen si un objeto pertenece a un subconjunto particular denominado *clase de decisión* o a un *concepto*. También vamos a suponer que la definición de esta clase es conocida porque por ejemplo lo ha definido el experto o un usuario. A partir de este instante, estamos ante un caso de los denominados de *aprendizaje supervisado*. Los objetos analizados se denominan *ejemplos de aprendizaje*. Las reglas objeto de estudio, como comprobaremos más adelante, son sentencias lógicas de la siguiente forma:

SI <condiciones> ENTONCES <clase de decisión>

Donde las condiciones se componen de una conjunción de tests elementales sobre los valores de los atributos.

Visto el planteamiento general y el objetivo de esta apartado, vamos a profundizar más en él, dado que la cuestión más importante del enfoque rough set es la obtención de las reglas de decisión del conocimiento  $K=\langle U, R \rangle$ .

Formalizando el problema: supongamos que los objetos de  $U$  en  $K$  se describen mediante un conjunto de atributos que a su vez se dividen en dos subconjuntos de atributos de condición y atributos de decisión, a los que vamos a denotar por  $C$  y  $D$ , respectivamente.  $D$  a menudo es un conjunto unitario (porque la decisión puede consistir en aceptar o rechazar, por ejemplo).

Es decir, es suficiente considerar tablas de decisión con una única decisión ya que siempre por medio de una simple codificación se puede transformar cualquier tabla de decisión con más de una decisión en una

tabla de decisión con exactamente una sola decisión. Se puede interpretar un atributo de decisión como un tipo de clasificación del universo de los objetos que ha sido formulada por un experto decisor, operador, físico... Observemos, como resultado, que el atributo de decisión  $d$  determina la partición del universo  $U$  en la denominada "clasificación de los objetos en  $U$  determinada por la decisión  $d$ ".

Por tanto, una tabla de información se transforma en una *tabla de decisión* si hay definido un conjunto  $C$  de *atributos de condición* y  $D$  de *atributos de decisión* (asignación a alguna clase), donde  $Q=C \cup D$  y  $C \cap D = \emptyset$ .

De una tabla de decisión  $S = \langle U, C \cup D, V, f \rangle$  se pueden derivar un *conjunto de reglas de decisión*. Asumamos que  $D$  es un conjunto unitario ( $D = \{d\}$ ) lo cual, como hemos mencionado, no disminuye la generalidad para consideraciones posteriores. Los conjuntos elementales  $d$  en  $S$  denotados por  $Y_j$ , ( $j= 1,2,\dots,n$ ), son llamadas *clases de decisión*. Si describimos las clases de decisión en términos de atributos de condición de  $C$  se consiguen las aproximaciones por encima y por debajo,  $\bar{C} Y_j$  y  $\underline{C} Y_j$ , respectivamente, además de la región frontera  $B_{n_c}(Y_j) = \underline{C} Y_j - \bar{C} Y_j$ ,  $j= 1,2,\dots,n$  de acuerdo con lo expuesto anteriormente.

Una tabla de decisión  $S = \langle U, C \cup D, V, f \rangle$  es *determinística* (consistente) cuando lo son todas las reglas que de ella se derivan, es decir, si  $C \rightarrow D$ . En otro caso, es no determinística (inconsistente). La tabla de decisión determinística describe *unívocamente* las decisiones (acciones) que deben tomarse cuando algunas *condiciones* se satisfacen. En el caso de una tabla no determinística, las decisiones no se determinan unívocamente por las condiciones (es decir, las condiciones pueden conducir a varias posibles decisiones). En su lugar, se define un conjunto de decisiones, el cual sería apropiado bajo circunstancias determinadas por condiciones. Estas últimas reglas son consecuencia de la información imperfecta de los objetos.

Una regla de decisión puede expresarse, tal y como hemos mencionado al principio de esta sección, como una sentencia lógica que relaciona la descripción de condiciones y las clases de decisión:

“**si** <conjunción de condiciones elementales> **entonces** <disjunción de decisiones elementales>”

Visto de una forma general que nuestra finalidad es buscar condiciones que si se satisfacen implicarán determinadas decisiones, vamos a formalizar este último objetivo.

Las fórmulas de condición elemental sobre el subconjunto  $A \subseteq C$  y el dominio  $V_{a_i}$  del atributo  $a_i \in A$  se definen como  $a_i = v_i$ , donde  $v_i \in V_{a_i}$ . Por  $\text{cond}_A$  denotamos una conjunción de fórmulas de condición elementales, esto es  $(a_i = v_i) \wedge \dots \wedge (a_r = v_r)$  para todo  $a_i \in A$ , y por  $[\text{cond}_A]$  entendemos el conjunto de todos los objetos que satisfacen la conjunción  $\text{cond}_A$ . Obviamente, si el objeto  $x \in [\text{cond}_A]$  entonces  $[\text{cond}_A] = I_A(x)$ .

De forma similar, definimos la formula de la decisión elemental  $d = v_j$ , donde  $v_j \in V_d$ . Por  $\text{dec}_D$  denotamos una disjunción de fórmulas de decisión elemental de la forma:  $(d = v_j) \vee \dots \vee (d = v_s)$  donde  $1 \leq s \leq n$ .

Es decir, el algoritmo de decisión tomaría la siguiente forma: *si* (atributo [1] = valor [i] y ... y (atributo [m] = valor [j] ) *entonces* (decisión [i] = clase [j]).

Sea  $J = \{ j: d = v_j \in \text{dec}_D \}$  un subconjunto de índices de clases de decisión  $Y_j$  representado en  $\text{dec}_D$ .

Por  $[\text{dec}_D]$  entendemos el conjunto de objetos, tanto de la aproximación por debajo  $\underline{C}Y_j$  de una clase de decisión  $Y_j$ , si  $s=1$ , o, en cualquier otro caso, de la frontera de varias clases de decisión  $Y_j$ , que satisfacen la disjunción  $\text{dec}_D$ . Precisando,

$[\text{dec}_D] = \underline{CY_j}$  para  $j \in J$  si  $J$  es un conjunto unitario (o sea, si  $s=1$ )

$[\text{dec}_D] = \bigcup_{j \in J} BN_c(Y_j)$  en cualquier otro caso.

La regla de decisión “**si**  $\text{cond}_A$  **entonces**  $\text{dec}_D$  “ es *consistente* si y solo si  $[\text{cond}_A] \subseteq [\text{dec}_D]$ . Si el conjunto  $J$  es unitario ( $s=1$ ), esto es que  $\text{dec}_D$  consta solamente de una decisión elemental, la regla de decisión es *exacta*, en otro caso es *aproximada*. Las reglas *aproximadas* son consecuencia de una descripción aproximada de las clases de decisión en términos de bloques de objetos (gránulos) no diferenciables por los atributos de condición. Esto significa que usando el conocimiento disponible, uno solo es capaz de decidir si algunos objetos (de la región frontera) pertenecen a una clase de decisión dada o no. En otras palabras, las reglas derivadas de utilizar la aproximación por debajo son exactas, mientras que las reglas derivadas son aproximadas cuando utilizamos la aproximación por encima o la región frontera.

El conjunto de reglas para todas las clases de decisión se denomina *algoritmo de decisión*.

Un algoritmo de decisión puede entenderse como la representación más compacta y sin redundancias correspondiente a un sistema de información. La primera condición se entiende en el sentido del menor número de reglas de decisión y el menor número de atributos que aparezcan en la definición de todas las reglas.

El algoritmo de decisión muestra todas las relaciones importantes utilizando un mínimo número de reglas y/o un mínimo número de atributos que aparezcan en las reglas. Por tanto, el algoritmo de decisión es más legible para el usuario que el sistema de información original.

Cada regla de decisión se caracteriza por su *fuerza*, esto es el número de objetos que satisfacen la parte de la condición de la regla (en

el argot se dice *cubiertos* por la regla) y que pertenecen a la clase de decisión sugerida. En el caso de reglas aproximadas, la fuerza se calcula para cada posible clase de decisión de forma separada. Las reglas más fuertes son normalmente más generales, es decir, su parte de condición es más corta y menos especializada.

Por tanto no todas las reglas de decisión del algoritmo de decisión son igualmente de importantes o de fiables para el decisor. Las reglas se crean sobre la base de diferentes números de objetos del sistema de información. Una regla puede crearse sobre la base de un único objeto mientras que otra puede crearse sobre la base de muchos objetos. En consecuencia, la primera regla es *más débil* que la segunda, o sea, menos *fiable* como ayuda a la hora de tomar una decisión. Esta es la razón por la que junto a la regla se indican el número de objetos que en el sistema de información han dado lugar a la misma, o sea, su fuerza.

La ingeniería del conocimiento ha dado lugar a un gran número de algoritmos para inducir generalización de los ejemplos. Si consideramos un conjunto de ejemplos de aprendizaje o casos, una generalización consiste en una descripción *discriminante* de cada clase o concepto. Una descripción de una regla de una clase de decisión se dice que es discriminante cuando es *completa* (cada ejemplo positivo, esto es, que pertenece a la clase, ha de ser reconocido como perteneciente a dicha clase de decisión) y *consistente* (cada ejemplo negativo, esto es que pertenece a cualquiera de las otras clases, no puede ser reconocido como perteneciente a la clase de decisión considerada anteriormente). La descripción discriminante es normalmente *mínima*, o sea, que si quitásemos cualquiera de sus reglas ya no resultaría una descripción completa.

Algunos algoritmos dan esas descripciones en forma de árboles de decisión, mientras que otros proveen reglas de decisión que son las que en este capítulo estamos considerando.

Los procedimientos para generar reglas de decisión a partir de una tabla de decisión operan sobre los principios del aprendizaje inductivo. Los objetos son considerados como ejemplos de decisiones. Para inducir reglas de decisión que describan un conjunto de objetos  $[dec_D]$ , los objetos que pertenecen a  $[dec_D]$  se denominan *positivos* y todos los demás *negativos*, según las definiciones vistas anteriormente.

En definitiva, una regla de decisión es *discriminante* si es completa y consistente, esto es, distingue los ejemplos positivos de los negativos, y si es mínima, esto es, que si quitando alguna condición elemental de  $cond_A$  (dando lugar a  $cond_{A'}$ ) resultaría que  $[cond_{A'}] \not\subseteq [dec_D]$  (violación de la consistencia de la regla).

Puede también resultar interesante buscar reglas *parcialmente discriminantes* (Mienko et al., 1996). Estas son reglas que cubrirían un número limitado de ejemplos negativos además de los positivos. Se caracterizan por un coeficiente denominado *nivel de discriminación*, que especificaría en qué medida la regla es consistente, esto es, cuál es el ratio de ejemplos positivos con respecto al total de ejemplos cubiertos por la regla.

Antes de entrar en los algoritmos de inducción de reglas existentes vamos a ver cuales serían las reglas que se derivarían del ejemplo de los candidatos que hemos ido desarrollando a lo largo de este capítulo.

Supongamos que de los tres reductos que teníamos el comité elige el segundo (en un caso real debería justificarse dicha elección). Por tanto el resto de criterios podría eliminarse y nos quedaría la siguiente tabla reducida:



<b>Criterios</b>	<b>C<sub>1</sub></b>	<b>C<sub>3</sub></b>	<b>C<sub>7</sub></b>	<b>D</b>
<b>Candidatos</b>				
<b>X<sub>1</sub></b>	4	4	1	A
<b>X<sub>2</sub></b>	3	4	1	R
<b>X<sub>3</sub></b>	3	3	2	R
<b>X<sub>4</sub></b>	5	5	2	A
<b>X<sub>5</sub></b>	4	5	1	A
<b>X<sub>6</sub></b>	3	3	3	R
<b>X<sub>7</sub></b>	4	5	2	A
<b>X<sub>8</sub></b>	4	4	2	A
<b>X<sub>9</sub></b>	4	4	2	R
<b>X<sub>10</sub></b>	5	5	2	A
<b>X<sub>11</sub></b>	5	4	2	A
<b>X<sub>12</sub></b>	5	4	2	A
<b>X<sub>13</sub></b>	4	3	2	R
<b>X<sub>14</sub></b>	3	4	3	R
<b>X<sub>15</sub></b>	4	5	1	A

De esta tabla se obtiene las siguientes reglas de decisión:

Regla 1: si  $c_1 = 5$  entonces  $d = A$

Regla 2: si  $c_3 = 5$  entonces  $d = A$

Regla 3: si  $c_1 = 4$  y  $c_7 = 1$  entonces  $d = A$

Regla 4 si  $c_1 = 4$  y  $c_3 = 4$  y  $c_7 = 2$  entonces  $d = A$  o  $R$

Regla 5: si  $c_1 = 3$  entonces  $d = R$

Regla 6: si  $c_3 = 3$  entonces  $d = R$ .

Hemos de destacar que cinco de las reglas son determinísticas y una de ellas no lo es. La regla no determinística es la 4 y es la que se deriva de los candidatos 8 y 9 que pertenecen a diferentes clases de

decisión. Define a una tercera categoría de decisión que incluiría los candidatos a los que se debería entrevistar.

Las reglas nos indican cual es la política que sigue el colegio al admitir o rechazar un candidato.

Visto el ejemplo continuemos con los algoritmos de decisión.

Los algoritmos de inducción existentes utilizan una de las siguientes estrategias:

- 1) generación de un conjunto mínimo de reglas que cubran todos los objetos de una tabla de decisión,
- 2) generación de un conjunto exhaustivo de reglas consistentes en todas las reglas posibles de una tabla de decisión,
- 3) generación de un conjunto de reglas de decisión "fuertes", incluso en parte discriminantes, cubriendo relativamente muchos objetos pero no necesariamente todos los objetos de la tabla de decisión.

Pueden establecerse dos perspectivas principales para la inducción de reglas de decisión derivadas de un conjunto de ejemplos según las finalidades que se persigan. Las más comunes se relacionan con:

- la inducción *orientada a la clasificación* y
- la inducción *orientada al descubrimiento*.

La inducción *orientada a la clasificación* tiene por objetivo el encontrar de manera automática, a partir de un conjunto de ejemplos de aprendizaje, un conjunto de reglas de clasificación que serán utilizadas para *construir* una clasificación.

El principal criterio para evaluar la calidad de un conjunto de reglas de clasificación es *el ratio de precisión de la clasificación* (tal y como se

ha definido en las secciones anteriores). Algunos estudios comparativos, basados en experimentos en los que se ha utilizado este criterio, han demostrado que la mayoría de los algoritmos dan buenos resultados y muy similares entre ellos. Hemos de resaltar que otros sistemas de clasificación, por ejemplo las redes neuronales o los métodos estadísticos dan también buenos resultados e incluso a veces mejores que los que dan los sistemas basados en reglas, pero estos últimos, a diferencia de los otros, dan una *representación simbólica del conocimiento* que encierra la clasificación que es menos intuitiva de interpretar.

El otro enfoque de inducción de reglas es la inducción *orientada al descubrimiento* y cuyo propósito es extraer, de un conjunto de datos, patrones de información y regularidades (reglas) “interesantes” y útiles para los diferentes tipos de usuarios. Tales reglas interesantes pueden ayudar a comprender y explicar las relaciones entre los valores de los atributos y las definiciones de las clases de decisión. El significado de “interesante” depende claramente del nivel de experiencia, las expectativas y de requisitos del usuario. En general, se consideran reglas interesantes las que se corresponden con patrones que cumplen estos tres requisitos: *fuertes* (medido a través del número de ejemplos positivos), *simples* (es decir, cuya parte de la condición consiste en un número bastante limitado de condiciones elementales fáciles de interpretar por el usuario; por tanto, se mide a través de la longitud de la regla) y *consistentes* (es decir, que la relación entre la parte de la condición y la parte de la decisión está suficientemente justificada). Este segundo enfoque es, por así decirlo, más general que el anterior al no centrarse únicamente en clasificar.

Dado que estos dos perspectivas persiguen objetivos distintos, los enfoques y algoritmos para cada uno de ellos también deben ser distintos. Sin embargo la mayoría de los algoritmos aplicados al segundo enfoque son los mismos que se utilizan para el enfoque de clasificación, lo cual provoca que den como resultado un *mínimo número de reglas* (objetivo

del enfoque de clasificación, para el que las reglas demasiado específicas no tienen importancia) que representan solo una parte limitada de las reglas “interesantes” que pueden existir. Es decir, pueden existir en el conjunto de datos otras reglas que no son interesantes desde el punto de vista de clasificación pero sí lo son en cuanto al conocimiento que de ellas se puede desprender.

Es necesario mencionar algunas de las diferencias entre uno y otro enfoque (además de las que ya hemos indicado previamente):

- en la inducción orientada al descubrimiento cada regla es evaluada de forma individual y de manera independiente, como una posible representante de un patrón específico; en el otro enfoque las reglas forman parte de un sistema y la evaluación se refiere al conjunto completo de reglas.

- en la inducción orientada al descubrimiento hemos indicado que las reglas deben ser fuertes, simples y consistentes, sin que ninguno de estos requisitos predomine sobre los demás; en la inducción orientada a la clasificación hay un criterio que domina y es el ratio de aproximación de la clasificación.

No vamos a entrar en todos los algoritmos desarrollados para la inducción de reglas de decisión basados en el enfoque rough set. Sin embargo, y con la finalidad de que la utilización del software informático ROSE en la parte empírica del trabajo que se basa en algunos de estos algoritmos, no se convierta en una especie de “caja negra”, describiremos brevemente algunos de ellos. Dado que nuestro objetivo va a ser clasificar, la mayoría de los algoritmos que describiremos son los que inducen reglas orientadas a clasificar, tan solo el primero que vamos a comentar induce reglas orientadas al descubrimiento. Solo vamos a entrar en una descripción general, sin profundizar en su formulación ni en los pasos que los desarrollan.

Stefanowski y Vanderpooten (1994) desarrollaron un algoritmo de inducción de reglas “interesantes” que posteriormente Mienko, *et al.* (1996) implantaron en un programa de ordenador. El punto de partida de este algoritmo es que la mayoría de los algoritmos orientados a la clasificación siguen una estrategia ambiciosa que consiste en crear una primera regla mediante la elección de manera secuencial de las “mejores” condiciones elementales de acuerdo a algún criterio de tipo heurístico. Posteriormente, los ejemplos de aprendizaje que encajan con esta regla no se tienen en cuenta. El procedimiento se repite mientras existan ejemplos significativos que todavía no se hayan descrito. Un punto crucial en estos algoritmos orientados a la clasificación es el modo en que desarrollan la búsqueda dentro del espacio de las reglas posibles para seleccionar “buenas” reglas y combinarlas en el subconjunto requerido. Dado que una caracterización explícita y una exploración exhaustiva de este espacio conduce a dificultades de tipo combinatorio, se suelen implantar *enfoques heurísticos* evitando estas caracterizaciones y pasos exploratorios. Todos estos algoritmos siguen un esquema que busca inducir *un mínimo conjunto de reglas* que cubra *todos* los ejemplos.

Para evitar estas dificultades Stefanowski y Vanderpooten (1994) sugirieron un enfoque alternativo en dos niveles que caracterizan y exploran, al menos en parte, el espacio de las reglas. Los dos niveles serían:

- 1- generación de todas las reglas “relevantes”
- 2- selección de algunas de ellas para construir una descripción “satisfactoria”.

Además a este algoritmo orientado al descubrimiento de reglas que sean fuertes, simples y consistentes, se le exigen otros dos requisitos: *el número de reglas* debe ser aceptable (es decir, no debería ser demasiado grande ya que entraría en conflicto con las capacidades cognitivas del usuario que son limitadas) y *la precisión de la clasificación* también debe

ser aceptable (comparándola con las tasas obtenidas con algoritmos orientados a la clasificación).

Sin embargo para obtener reglas que sean fuertes, simples y consistentes y que sean “interesantes” desde el punto de vista del usuario debemos definir los niveles de *fuerza*, *longitud* y *grado de discriminación* en las reglas que queremos generar. Las reglas obtenidas se deben comparar con los otros dos requisitos mencionados y teniendo en cuenta en la definición de los niveles, una vez más, los requisitos del usuario. No hay que olvidar que dichos parámetros dependerán de cada conjunto de datos.

La exploración del espacio de las reglas que realiza el algoritmo se repite de manera iterativa para cada concepto que ha de describirse. El algoritmo genera de manera progresiva reglas de tamaño cada vez mayor comenzando por las más cortas. La estrategia comienza con una regla inicial que tiene la parte de la condición vacía. Durante el proceso de búsqueda el conjunto vacío se va completando con condiciones elementales dentro de una lista de las permitidas. Estos conjuntos se van evaluando como candidatos a integrar la parte de la condición de las reglas.

Sin embargo el primer nivel, *de generación de todas las reglas*, tal y como sugieren los propios autores, no debe utilizarse tal cual, sin depurar previamente, ya que conduciría a una gran complejidad tanto de tiempo como de cálculo. Buscamos *todas las reglas relevantes*. Por tanto, la exploración del espacio de las reglas se controla mediante la incorporación de una o varias *condiciones de detención* (que están relacionadas con los requisitos del usuario y en concreto, con la fuerza y simplicidad de las reglas) y que conducen la búsqueda hacia las reglas *relevantes* (los patrones de conocimiento más fuertes). Las condiciones de detención garantizan las propiedades que son deseables para estas reglas y reducen de manera significativa la carga computacional del algoritmo. Al introducir estas condiciones puede ocurrir que algunos

ejemplos no queden cubiertos por la regla de decisión. Esto lejos de perjudicar, puede ser instructivo ya que pueden presentarse al usuario como ejemplos atípicos o, si no lo son, indicarían que a lo mejor hay que relajar las condiciones de detención.

Respecto al segundo nivel, de construcción de una *descripción satisfactoria*, implica que no se busca una descripción óptima (lo que sería en pureza el objetivo, de entre todas las reglas encontrar las mejores) sino una solución suficientemente próxima a la óptima (establecida de acuerdo a algún criterio de preferencias como puede ser, por ejemplo, que tenga un mínimo número de condiciones elementales en función de la complejidad de la descripción, o minimizar el número de reglas). Por lo que algunos requisitos se pueden transgredir siempre y cuando dicha trasgresión esté controlada y evitada si así fuese requerido. Esto también disminuye el tiempo de cómputo.

Finalmente, debemos comentar que este algoritmo puede adaptarse fácilmente para manejar datos tanto incompletos como inconsistentes y en consecuencia utiliza los conceptos del enfoque rough set para este tipo de datos.

El primero de los algoritmos de inducción de reglas orientadas a la clasificación que vamos a describir es el desarrollado por Grzymala-Busse (1992), el LERS. El sistema ROSE, aplicación que utilizaremos en la parte práctica de la investigación, genera reglas de decisión utilizando una versión modificada de este algoritmo (introducida por el autor antes mencionado), el LEM2. Este algoritmo se compone, a su vez, de un grupo de algoritmos de inducción, que están enfocados a inducir tanto la denominada *descripción discriminante* de cada clase de decisión como una descripción muy cercana a ella.

Muchos de los algoritmos de inducción de reglas de decisión han dado lugar a aplicaciones informáticas que pueden incorporar uno o varios algoritmos. En esta situación se encuentra el sistema LERS. Su

aplicación informática ha sido tan utilizada que, tal y como hemos indicado, incluso ha sido incorporado al ROSE, por lo que vamos a dejar su descripción para el capítulo 3 dedicado al software utilizado en la parte empírica del trabajo y, en concreto, en los antecedentes del mismo. Aquí nos hemos limitado a indicar en qué consiste y cuál es su objetivo.

Skowron (1993) propuso un método para la generación de reglas de decisión óptimas con coeficientes ciertos. Además puede aplicarse a la generación de reglas de decisión con un número mínimo de descriptores en la parte de la disjunción, o sea, en el lado izquierdo de las reglas.

Las reglas de decisión que considera son de la forma siguiente:  $\Gamma \Rightarrow \Gamma'$  donde  $\Gamma$  y  $\Gamma'$  son combinaciones booleanas de descriptores sacadas de las condiciones y decisiones de un experto, respectivamente. Dichas reglas se generan con algunos coeficientes ciertos expresados mediante funciones básicas de la teoría de la evidencia y funciones de pertenencia rough calculadas de una tabla de decisión dada. Estos coeficientes se pueden utilizar en la toma de decisiones.

El método de la generación de reglas se basa en la construcción de funciones booleanas apropiadas derivadas de las matrices de diferenciación modificadas (Skowron y Rauszer, 1992).

Se consideran, además, dos clases de reglas de decisión óptimas: reglas óptimas localmente y reglas óptimas globalmente. El método también se puede aplicar a la construcción de reglas de decisión basadas en las aproximaciones por arriba y por debajo.



## VIII.- AYUDA A LA TOMA DE DECISIONES UTILIZANDO REGLAS DE DECISIÓN

Las reglas de decisión derivadas de una tabla con el enfoque rough set se expresan en términos de los atributos más significativos sin las redundancias típicas de los datos originales y, en consecuencia, pueden ser utilizadas para recomendaciones sobre la clasificación de nuevos objetos.

Un *problema de clasificación*, es un tipo de problema de decisión, tal y como veremos en la sección 9 de este capítulo, que consiste en determinar la asignación de un objeto, descrito por un conjunto de atributos, a una clase de decisión conocida a priori. La ayuda a la clasificación se basa en el uso de reglas de decisión derivadas de la experiencia pasada (manifestada a través de un conjunto de ejemplos de aprendizaje) y sistematizada en forma de reglas a través del enfoque rough set.

La clasificación de cualquier *nuevo objeto* puede apoyarse comparando su descripción con cada una de las reglas de decisión obtenidas, es decir, comparando dicho objeto con las condiciones de las reglas de decisión.

Esta comparación puede conducir a una de las siguientes situaciones (Slowinski, 1993 y Slowinski y Stefanowski, 1994):

- a) el nuevo objeto encaja exactamente con una regla de clasificación determinística (exacta),
- b) el nuevo objeto encaja con más de una regla determinística (exacta) indicando, sin embargo la misma clase de decisión,
- c) el nuevo objeto encaja con una regla aproximada (no determinística) o más de una regla aproximada que indican diferentes

clases de decisión, o más de una regla determinística que también conduzcan a diferentes clases de decisión,

d) el nuevo objeto no encaja con ninguna de las reglas.

En los casos a) y b), la recomendación es unívoca mientras que en el caso c) de comparación no clara o ambigua, el usuario es informado de la *fuerza* total (el número de objetos clasificados que soportan cada categoría posible, en otras palabras, que verifican cada posible decisión indicada por la regla) de todas las reglas de decisión con las que encaja con respecto a las clases de decisión sugeridas. Si la fuerza de una categoría es mayor (esto es, el número de objetos es mayor que el 51% de todos los objetos que han dado lugar a dicha regla) que la fuerza de otras categorías dentro de las que existen en la regla no-determinística, se puede concluir que de acuerdo a esta regla, el objeto considerado lo más probable es que pertenezca a la categoría *más fuerte*. La información acerca de la fuerza de cada regla puede darse de manera extendida presentando al decisor los objetos que soportan un regla dada.

Si ninguna regla de decisión encaja con el nuevo objeto (caso d), el decisor puede conseguir el conjunto de reglas "más cercanas" a la descripción del objeto. Las reglas más cercanas son reglas que están próximas a la descripción de un objeto clasificado en el sentido de una determinada medida de distancia (en otras palabras, no difieren del objeto de un modo significativo). El motivo de introducir las reglas "más cercanas" resulta de la creencia de que ofrecer al decisor más información sobre la vecindad de un objeto clasificado puede darle alguna sugerencia en lugar de no ofrecerle nada.

La noción de *proximidad* implica, tal y como se ha indicado, el uso de una medida de distancia. Aunque el software que hemos utilizado en la parte empírica del trabajo (ROSE) incorpora la medida de distancia denominada métrica  $L_p$  (Stefanowski 1993 a y b, Slowinski y Stefanowski 1992), vamos a comentar la que incorpora el RoughClass, software

predecesor del ROSE, y sobre la cual se han basado múltiples aplicaciones. Esta medida es la *Relación de Proximidad Valorada*.

Antes de presentarlas, vamos a comentar brevemente la naturaleza real de los atributos analizados, dado que la definición de ambas se verá afectada según cual sea la naturaleza de los atributos.

En general, hay que tener en cuenta a la hora de comparar los objetos con las reglas dos clases de atributos: cardinales y nominales (aunque existen más tipos de atributos, distinguiremos sólo estos dos tipos y con las definiciones que exponemos a continuación).

Los atributos cardinales (también ordinales) son aquellos cuyo dominio es ordenado de acuerdo a una escala numérica. Por tanto, sus valores pueden ser ordenados de mayor a menor. Son de carácter *cuantitativo*. El dominio de los atributos nominales puede no ser ordenado dado que consisten en códigos lingüísticos (ordenados o no) o lo que es lo mismo son de carácter *cualitativo*. En consecuencia, cuando hablamos de atributos nominales asumimos que no se ordenan. Por tanto, para un atributo cardinal se puede calcular una diferencia entre dos valores mientras que para los nominales es posible únicamente revisar si existe una diferencia entre sus valores o no.

En la realidad a menudo coexisten ambos tipos de atributos. Así por ejemplo en las aplicaciones prácticas en el campo de la medicina el conjunto de atributos habitualmente consisten en una mezcla tanto de tipo cuantitativo como de tipo cualitativo.

A menudo, incluso, los valores de los atributos cuantitativos no son interpretados por los físicos como números precisos, sino que se comparan con alguna norma establecida. Por ejemplo, se explican a un nivel bajo, medio o alto. En otras palabras, los atributos cuantitativos son, como veremos, a menudo traducidos para interpretaciones prácticas en términos cualitativos.

Además el problema de encontrar las reglas más cercanas se relaciona con otras dos cuestiones: la posibilidad de un tipo de inconsistencia del algoritmo de decisión y de grados distintos de fiabilidad de las reglas.

El primero es consecuencia de la *hipótesis de un mundo cerrado*, esto es, se supone que la experiencia contenida en el sistema de información es completa. Es necesario resaltar que si un objeto no encaja con ninguna regla, significa que no existe ningún caso en la experiencia representada en el sistema de información. Este problema es consecuencia del hecho de que el análisis rough set de la experiencia representada en un sistema de información se desarrolla bajo la *hipótesis de un mundo cerrado*, mientras que la clasificación de nuevos objetos se realiza en la práctica bajo la *hipótesis de un mundo abierto*. Además, la aplicación de un algoritmo de decisión obtenido bajo la *hipótesis de un mundo cerrado* para clasificar objetos que provienen de "un mundo abierto" puede llevar a un tipo de inconsistencia dado que estos algoritmos funcionan correctamente con objetos cuya descripción es consistente con la experiencia previa. Pero para objetos en los que esto no ocurre puede dar lugar a clasificaciones no unívocas de un objeto por medio de reglas determinísticas que encajan exactamente con la descripción del mismo. Es decir, un objeto puede encajar con dos reglas determinísticas que indiquen diferentes decisiones. Este es un caso típico de la *hipótesis de un mundo abierto*, en la que la experiencia representada en un sistema de información no es completa.

Por esta razón, es necesario antes de utilizar un algoritmo para clasificar nuevos objetos (si este es nuestro problema de decisión) contrastar la consistencia del mismo.

El segundo punto está relacionado con el hecho de que no todas las reglas son igual de importantes para el decisor. Algunas reglas puede incluir muchos objetos mientras que otras solo unos pocos o incluso uno único. Por tanto tienen diferente *fuerza*. La introducción de este

coeficiente es útil para el análisis de las reglas no determinísticas. En este tipo de reglas las decisiones no están definidas unívocamente por las condiciones. De este modo, si definimos la fuerza para cada posible clase, podemos concluir que si la fuerza de una clase es mucho mayor que la de otra, lo más probable es que el objeto que pretendemos clasificar pertenezca a la clase más fuerte.

Otra cuestión a considerar es que puede haber menos condiciones en la regla que la descripción que tenemos del objeto que vamos a clasificar, por tanto, la distancia entre ellos (para cualquiera de las dos medidas que vamos a presentar) solo se calcula únicamente para los atributos representados en la regla. Para el resto de atributos se considera que no existe diferencia.

### **Métrica $L_p$**

Por  $A$  denotamos el conjunto de atributos  $(a_1, a_2, \dots, a_n)$  Los atributos ordinales los denotaremos por  $O$  y los nominales por  $N$ . Por tanto,  $A = O \cup N$ .

Consideremos un objeto nuevo  $x$  cuya descripción es  $(x_1, x_2, \dots, x_m)$  siendo  $m \leq n$  ya que únicamente vamos a considerar los atributos que hay en la parte de condiciones de la regla  $y$ .

La distancia  $x$  a la regla  $y$  (descrita por los valores de los atributos  $(y_1, y_2, \dots, y_m)$ ) se define como:

$$L_p = \frac{1}{m} \left( \sum_{i \in N} k_i d_i^p + \sum_{j \in O} k_j d_j^p \right)^{1/p}$$

Donde:

$m$  - número de atributos

$p$  - es un parámetro (número natural) elegido por el analista para el problema considerado,  $p = 1, 2, \dots$

$d_i$  - es una distancia parcial para atributos nominales,  $d_i = 0$  si  $x_i = y_i$  = 0 y  $d_i = 0.5$  en cualquier otro caso

$d_j$  - es una distancia parcial para atributos ordinales,  $d_j = |x_i - y_i| / (v_{jmax} - v_{jmin})$  siendo  $v_{jmax}$  y  $v_{jmin}$  los valores máximo y mínimo respectivamente del atributo  $a_j$

$k_i$  y  $k_j$  - son los pesos (coeficientes de importancia) de los atributos.

Los pesos pueden determinarse tanto de manera subjetiva como tomando en consideración la capacidad clasificatoria del atributo (por ejemplo, la capacidad clasificatoria del atributo  $a_i$  puede expresarse por la diferencia entre la calidad de clasificación del conjunto de todos los atributos considerados y la calidad de clasificación para el mismo conjunto de atributos pero sin el atributo  $a_i$ ).

La distancia se calcula para cada regla y estas se ordenan en función del valor que tome  $L_p$ . Las reglas con menores valores son las que se presentan al decisor.

Es necesario mencionar que dependiendo del valor que tome  $p$ , la medida de distancia es más o menos compensatoria. Por ejemplo, valores pequeños de  $p$  permiten que una gran diferencia en un criterio sea compensada con un número de pequeñas diferencias en otros criterios. Para valores altos de  $p$ , se produce una tendencia a sobrevalorar la mayor de las diferencias parciales hasta el punto de ignorar todas las otras diferencias parciales para  $p = \infty$ .

Esto es un punto débil a tener en cuenta ya que la compensación posibilita que una diferencia grande en un atributo sea compensada por pequeñas diferencias en otros atributos.

Otra medida de distancia se puede obtener cuando el decisor ha definido para cada atributo una restricción en el sentido de permitir hasta una distancia determinada. Para un atributo ordinal definiremos una

distancia máxima aceptada, por tanto, si la diferencia entre la regla y el objeto es mayor que el nivel permitido, la regla no puede ser tratada como cercana. Para un atributo nominal, el decisor debe definir pares de valores de los atributos que en su opinión son demasiado distantes (es decir, prohibitivos y si esas relaciones se dan en el cálculo de la distancia a la regla, ésta no puede considerarse cercana). Estas nuevas restricciones eliminan también las grandes diferencias entre el objeto clasificado y la regla. Como resultado puede usarse como criterio el número de diferencias en lugar de la medida agregada de las cantidades de las diferencias. Por tanto, la nueva medida de distancia sería:

$$D' = d/m$$

Donde  $m$  es el número de atributos y  $d$  es el número de condiciones en la regla que son diferentes de la descripción del objeto.

Las reglas más cercanas se pueden presentar al decisor con la siguiente información: posibles asignaciones de un objeto sobre la base de una regla dada, fuerza de la regla, objetos que la crean, número y posición de las diferencias.

También hay que mencionar que centrarse en el número de diferencias en lugar de agregar la información sobre su cantidad también puede ser considerado como un punto débil o criticable.

### **La Relación de Proximidad Valorada (RPV)**

Slowinski (1993) propuso una medida de distancia basada en una *relación de proximidad valorada*, RPV.

La relación de proximidad valorada evita que una diferencia mayor en un atributo sea compensada por un número de pequeñas diferencias en otros atributos.

Ya que una regla de decisión puede tener menos condiciones elementales (criterios) que la descripción de un objeto que ha de ser

clasificado, la proximidad (distancia) de la regla de clasificación al objeto se calcula solamente para los atributos representados. Las diferencias entre el objeto y la regla en los otros atributos no se toman en consideración. Por tanto, se asume que no existe diferencia para otros atributos (criterios).

La *relación de proximidad valorada*  $R$  se construye utilizando un principio similar al de la *relación de sobreclasificación valorada* introducida por Roy (1985), haciendo uso de los niveles de preferencia, indiferencia y veto de atributos particulares, mediante la utilización de los tests de concordancia y discordancia similares a los empleados para la construcción de la relación de sobreclasificación por Roy (1985). Los objetivos de estos tests son:

- caracterizar a un grupo de atributos considerados que están en concordancia con la afirmación “*el objeto  $x$  está próximo a la regla  $y$* ”, y valorar la importancia relativa de este grupo,
- caracterizar entre los atributos que no están en concordancia con la afirmación de arriba, aquellos cuya oposición es lo suficientemente fuerte como para reducir la credibilidad de la cercanía, que sería lo que resultaría de tener en consideración sólo la concordancia, y calcular la posible reducción que resultaría de este modo.

Por tanto, dado un nuevo objeto  $x$  descrito por los valores  $a_1(x)$ ,  $a_2(x)$ , ...,  $a_m(x)$  correspondientes a los atributos representados en la parte de la condición de la regla  $r$  ( $m \leq \text{card}(C)$ ), es decir, la regla de decisión tiene menos condiciones/atributos que  $n$ . Asumamos que todos los atributos son cuantitativos. El objeto  $x$  se compara con cada regla de clasificación  $r$  para valorar la credibilidad de la relación de proximidad  $x(RPV)r$  (“ $x$  es próximo a  $r$ ”). El cálculo de la credibilidad  $\delta(x,R)$  de la proximidad se basa en el sentido común: la formula para determinar  $\delta(x,R)$  en el intervalo  $[0,1]$  se construye para que respete ciertos principios cualitativos, y, en particular, para que excluya la posibilidad de



compensaciones indeseadas. La credibilidad  $\delta(x,R)$  es igual a 1 si la afirmación  $x(RPV)r$  está bien fundada;  $\delta(x,R)$  es igual a 0 si no hay argumentos para la proximidad de  $x$  y  $r$ .

La fórmula para el cálculo de  $\delta(x,R)$  se basa en dos conceptos llamados *concordancia* y *discordancia* (Roy y Bouyssou, 1993).

Para realizar los cálculos, el usuario tiene que expresar explícitamente y numéricamente:

- la relativa importancia (peso)  $k_i$  que el usuario desea conferir al atributo  $a_i$  en el cálculo de la concordancia,
- el mínimo valor de la discordancia que da al atributo  $a_i$  el poder para rechazar la credibilidad de la afirmación de proximidad, incluso si se opone a todos los otros atributos en concordancia con la afirmación; esto se denota por  $v_i[a_i(x)]$  y se llama *el nivel de veto* del atributo  $a_i$ .

El índice de concordancia comprensivo se define como:

$$C(x,r) = \frac{\sum_{l=1}^m k_l c_l(x,r)}{\sum_{l=1}^m k_l}$$

Donde  $c_l(x,r)$  es un índice de concordancia parcial para el atributo (condición)  $a_l$ .

Para *atributos cardinales*, el cálculo de  $c_l(x,r)$  implica explicitar por parte del decisor dos niveles:  $0 \leq q_l[a_l(x)] \leq p_l[a_l(x)]$ , llamados niveles de *indiferencia* y *estricta diferencia*, respectivamente. Los pesos y los niveles definidos arriba vienen del conocimiento del dominio del usuario (experto, decisor).

Los niveles tiene el siguiente significado:

- si  $|a_i(r) - r_i| \leq q_i[a_i(x)]$ , el decisor no ve ninguna diferencia entre  $x$  y  $r$  para el atributo  $a_i(x)$ ,
- si  $q_i[a_i(x)] < |a_i(r) - r_i| < p_i[a_i(x)]$ , el decisor percibe una pequeña diferencia entre  $x$  y  $r$  para el atributo  $a_i(x)$ ,
- si  $|a_i(r) - r_i| \geq p_i[a_i(x)]$ , el decisor percibe una gran diferencia entre  $x$  y  $r$  para el atributo  $a_i(x)$ .

Para cada atributo que se opone a la afirmación que  $x(RPV)r$ , se calcula el índice de discordancia,  $d_i(x,r)$ , y tal y como se ha indicado anteriormente, para el caso de atributos cardinales el decisor ha de expresar explícitamente y numéricamente el nivel de veto  $v_i[a_i(x)]$ ; siendo  $v_i[a_i(x)] \geq p_i[a_i(x)]$ . Es el mínimo valor de discordancia que da al atributo  $a_i(x)$  el poder de descartar la credibilidad de la afirmación de la proximidad, incluso si se opone a todos los otros atributos.

La curva del índice de concordancia  $c_i(x, r)$  y del índice de discordancia  $d_i(x, r)$  se muestra en el siguiente gráfico para atributos cardinales. Resaltemos que la escala del eje  $Y$  está de acuerdo a la definición del atributo cardinal. Los atributos cardinales se suelen discretizar en el enfoque rough set.

Para *atributos nominales*, el decisor no puede utilizar los niveles para evaluar diferencias parciales. En su lugar, tiene que especificar para todos los pares de valores de atributos, si son indiferentes, ligeramente diferentes, estrictamente diferentes o excesivamente diferentes. Estas definiciones corresponden respectivamente a los intervalos definidos por los niveles  $q_i, p_i, v_i$ , sobre la escala de atributos cardinales (esto es,  $[0, q_i]$ ,  $(q_i, p_i)$ ,  $[p_i, v_i)$ ,  $[v_i, \infty)$ ).

La definición de los índices de concordancia y discordancia para atributos nominales es la siguiente:

- $c_i(x,r)=1$  y  $d_i(x,r) = 0$ , para un par atributo-valor indiferente,

- $c_i(x,r)=0,5$  y  $d_i(x,r) = 0$ , para un par atributo-valor ligeramente diferente,
- $c_i(x,r)=0$  y  $d_i(x,r) = 0,5$ , para un par atributo-valor estrictamente diferente,
- $c_i(x,r)=0$  y  $d_i(x,r) = 1$ , para un excesivamente diferente par atributo-valor.

Si el decisor se “resiste” a ciertos pares de valores de atributos, se asume que son estrictamente diferentes.

El grado de credibilidad  $\delta(x, r)$  de la relación de proximidad  $x(RPV)r$  se obtiene del índice de concordancia (índice de concordancia global) ponderado por los índices de discordancia (hasta el punto de su anulación):

$$\delta(x,r) = C(x,r) \prod_{i \in L} \frac{1 - d_i(x, r)}{1 - C(x, r)}, \quad L = \{i: d_i(x,r) > C(x,r)\}$$

El usuario puede decidir de acuerdo a la información sobre las reglas con los mayores valores de  $\delta(x, r)$  y la fuerza de las clases que son sugeridas.

Es decir, las reglas  $r$  con los mayores valores de  $\delta(x,R)$  se presentan al decisor junto con la información relativa a la fuerza de las correspondientes categorías (decisiones).

Consideremos el siguiente ejemplo:

Supongamos que tenemos un paciente al que se le han observado los siguientes síntomas y valores para cada uno de ellos:

Síntoma 1 - nivel de transaminasas,  $S_1 =$  alto

Síntoma 2 - fiebre,  $S_2 =$  baja

Síntoma 3 - dolor abdominal,  $S_3 =$  grande

Síntoma 4 - antecedentes,  $S_4 =$  alcohólico

Síntoma 5 - peristaltismo,  $S_5 =$  sí

Síntoma 6 - edad,  $S_6 =$  anciano

Hay un único atributo de decisión  $d$  que nos valora la gravedad de la enfermedad: A, grave o B, no grave.

Supongamos que del sistema de información se han obtenido únicamente estas dos reglas:

Regla nº 1 ( $r_1$ ) si  $S_1 =$  medio y  $S_3 =$  medio entonces  $d = A$

Regla nº 2 ( $r_2$ ) si  $S_1 =$  medio y  $S_2 =$  baja y  $S_3 =$  pequeño entonces  $d = B$

Como podemos observar el paciente no encaja con ninguna de las dos reglas por lo que vamos a comprobar cual es la regla más cercana, para ello hemos de especificar para cada síntoma los niveles de indiferencia, estricta diferencia y veto, además de los pesos que atribuidos a cada síntoma. Supongamos que son los siguientes (sólo consideramos los síntomas que nos afectan):

	$q_i$	$p_i$	$v_i$	$k_i$
$S_1$	0	1	2	2
$S_2$	0	0	2	2
$S_3$	0	0	2	2

Regla 1:

- el índice de concordancia parcial es  $c_3(x, r_1) = 1$ , por lo que el índice de concordancia global tomará el siguiente valor:

$$C = \frac{k_3 c_3}{k_1 + k_3} = 0.5$$

- para calcular el índice de discordancia para el atributo 1, lo primero que haremos es verificar que la diferencia entre el valor del atributo en el paciente y en la regla no supere o iguale el valor del nivel veto, cosa que sucede ya que la diferencia es uno al hacer una hipotética escala de la siguiente manera:

bajo -1 medio -2 alto- 3 muy alto - 4

$$d_1(x, r_1) = \frac{\text{diferencia en la escala}}{\text{amplitud de la escala}} = \frac{1}{3} = 0.33$$

- como el índice de concordancia es mayor que el de discordancia, el grado de credibilidad  $\delta(x, r)$  de la relación de proximidad  $x(RPV)r$  de la regla viene dada únicamente por el índice de concordancia global:

$$\delta(x, r) = 0.5$$

Regla 2:

Con la regla número dos procedemos de forma similar.

- el índice de concordancia parcial es  $c_2 = 1$ , por lo que el índice de concordancia global tomará el siguiente valor:

$$C = \frac{k_2 c_2}{k_1 + k_2 + k_3} = 0.33$$

- para calcular el índice de discordancia, lo primero que observamos es que la diferencia entre los valores de los atributos 1 y 3 en el paciente y en la regla iguala en ambos casos valor del nivel veto (2). Por lo que  $\delta(x, r)$  toma el valor cero, al ser los índices de discordancia iguales a 1.

Finalmente, es necesario mencionar que en la situación c) aparecería otra vez la ambigüedad. Entonces, la sugerencia puede basarse tanto en la fuerza de las posibles categorías, o en un análisis de los ejemplos de clasificación sobre los que se sustenta cada posible

categoría. En el último caso, la categoría sugerida es aquella que se apoya por un ejemplo de clasificación que sea el más cercano al nuevo objeto, en el sentido de la relación  $R$ .

## IX.- TIPOS DE PROBLEMAS DE DECISIÓN

Los problemas de clasificación son el punto central para la metodología rough set.

Una vez que hemos obtenido el algoritmo de las reglas de decisión, el cual representa el conocimiento que hemos obtenido sobre los casos que teníamos en un sistema de información, sería interesante y deseable utilizar este conocimiento para justificar la clasificación de nuevos objetos, tema que ya hemos comentado en la sección previa.

Por nuevos objetos entendemos aquellos que no estén contemplados en el sistema de información del experto y que están descritos únicamente por valores de atributos de condición. La asignación de estos objetos a clases es desconocida y el objetivo del experto es descubrir su asignación sobre la base del conocimiento de su experiencia previa.

Por tanto, se trata, en definitiva, de encontrar en el algoritmo de decisión aquella regla cuya parte de la condición coincida con la descripción del nuevo objeto. Si se encuentra dicha regla la decisión es clara y si no se encuentra, tal y como ya hemos indicado, se buscan las reglas más cercanas.

Asimismo, existen otros problemas de decisión en los que se puede aplicar el enfoque rough set (además del de clasificación). Dado que nuestro objetivo será clasificar empresas nos detendremos más en este último tipo de problemas, pero queremos comentar las otras posibilidades que ofrece la aplicación de esta teoría.

Dado un sistema de información  $S$  en el que un conjunto finito de objetos  $U$  se describen mediante un conjunto finito  $Q$  de atributos multivalorados, distinguimos tres clases de problemas de decisión:

- a) problemas de clasificación con atributos múltiples

- b) problemas de clasificación múltiple con atributos múltiples
- c) descripción de objetos con atributos múltiples.

Aunque esta clasificación de los problemas de decisión se solapa con la clasificación efectuada por Roy (1985) donde también se diferencian los problemas de clasificación con criterios múltiples y los problemas de descripción, sin embargo Pawlak y Slowinski (1994) difieren en la significación de una descripción con atributos múltiples. Esto demuestra que el análisis multicriterio y el análisis rough set representan filosofías diferentes de analizar problemas de decisión con atributos múltiples.

Una característica clave en la diferenciación de los problemas a) y b), y el problema de descripción c) es la división del conjunto de atributos  $Q$  en los subconjuntos de atributos  $C$  de condición y el subconjunto de atributos  $D$  de decisión que está presente en a) y b) pero no en c). Aún más, en los problemas de clasificación, uno o varios agentes están explícitamente involucrados en la situación de decisión y están representados mediante los atributos de decisión. En la descripción con atributos múltiples de objetos, dependiendo de la interpretación del sistema de información, los agentes pueden estar representados por los objetos o por los atributos, o pueden no estar representados ni por los objetos ni por los atributos y permanecer “alejados” de la descripción. Otra diferencia entre los problemas a) y b), y los problemas de descripción c), es que la principal cuestión relacionada con los primeros es la recomendación o prescripción mientras que con los últimos es la explicación, de acuerdo con lo que hemos expuesto en la introducción de este capítulo.

El problema a) es un problema clásico de clasificación con atributos múltiples donde sólo existe un atributo de decisión. Consiste en asignar cada objeto a una categoría apropiada previamente definida, por ejemplo: rechazo, aceptación o requerimiento de información adicional. La



información de las preferencias adquirida de un agente con respecto a este problema consiste en un conjunto de ejemplos de clasificación presentados en una tabla de decisión, y la metodología rough set se utilizará para analizar dicha información. Los ejemplos de clasificación pueden ser prototipos para enseñar ya sean construidos a propósito o situaciones reales, u observaciones efectuadas por un agente en el pasado.

Los problemas b) y a) difieren entre sí en la existencia de múltiples atributos de decisión. Esto podría significar, por ejemplo, que el mismo conjunto de objetos ha sido clasificado por varios agentes.

Tal y como hemos mencionado previamente, una más amplia distinción de problemas de decisión y descripción se puede efectuar sobre la base de la interpretación del sistema de información. Más específicamente, si los agentes están explícitamente representados ya sea por los objetos como por los atributos, distinguimos:

c.1.) Descripción con atributos múltiples de una situación de decisión: En este caso, los valores de los atributos representan opiniones de los agentes sobre cuestiones específicas de una situación de decisión. Si los agentes no están explícitamente representados en el sistema de información pero existen “alejados” de la descripción, nos interesa básicamente el siguiente apartado:

c.2.) descubrir las dependencias entre los atributos: Los atributos se interpretan como consecuencias de las decisiones que representan los objetos.

Vamos a especificar un poco más cada uno de los problemas considerados, teniendo en cuenta que todas las consideraciones que hagamos se refieren a la metodología rough set y que pueden no ser extrapolables a otros enfoques.

a) Problema de clasificación con atributos múltiples

Una vez que tenemos un conjunto de objetos descrito por un número de atributos, el problema de clasificación consiste en la asignación de cada objeto de un conjunto a una categoría previamente definida apropiada (por ejemplo: aceptar, rechazar o requerir información adicional).

Cuando los objetos se evalúan desde múltiples puntos de vista (atributos), el problema no tiene solución a menos que se proporcione información adicional acerca de las preferencias del decisor. Teniendo la información sobre las preferencias, se pueden agregar los múltiples puntos de vista en un modelo de preferencia global que conduzca a la solución del problema de clasificación con atributos múltiples.

Muy frecuentemente, el modelo de preferencia global tiene la forma de una función real (por ejemplo, la función de utilidad) o de relación (relación de sobreclasificación, Roy (1985)). Sin embargo, puede también consistir, tal y como hemos visto, en un conjunto de sentencias lógicas del tipo “ si...entonces...” denominadas reglas de clasificación que describen las preferencias del decisor.

En este último caso, el conjunto de reglas se derivan de la información sobre las preferencias expresadas *implícitamente* a través de ejemplos de decisiones de clasificación tomadas por el decisor (experto) sobre un conjunto de objetos (muestra de prueba).

En este caso, el problema de clasificación de un nuevo objeto en base a múltiples atributos se efectúa comparando su descripción con una de las reglas de clasificación; si esto “falla”, se le presenta al decisor el conjunto de reglas más cercanas, según lo indicado anteriormente.

En nuestro caso en concreto, si consideramos que la información acerca de las preferencias que viene del decisor, o del experto, está dada en forma de una tabla de decisión donde los objetos corresponden a

ejemplos, los atributos de condición a los criterios y el único atributo de decisión consiste en las decisiones sobre la asignación a una categoría, las filas de la tabla serían los ejemplos de las decisiones de clasificación relativas a los correspondientes objetos.

Esperamos los siguientes resultados del análisis rough set de la información de las preferencias:

- evaluación de la importancia de los atributos de condición para la definición de la clasificación de los objetos mediante valores de los atributos de decisión elegidos,
- construcción del menor subconjunto de criterios independientes que aseguren la misma calidad de clasificación que el conjunto total, esto es, los reductos del conjunto de atributos, mediante la reducción tanto de los objetos como atributos redundantes,
- si la intersección de los reductos no es un conjunto vacío nos dará un núcleo formado por los criterios que no pueden eliminarse sin menoscabar la capacidad de aproximar las reglas de decisión,
- eliminación de los criterios redundantes de la tabla de decisión,
- generación de las reglas de clasificación de la tabla de decisión reducida; estas incluyen únicamente los criterios relevantes y explicarían la política de toma de decisiones del decisor (experto).

Por supuesto, el resultado más importante desde el punto de vista de la ayuda en el proceso de toma de decisiones es el conjunto de reglas de clasificación. Constituye un modelo global de las preferencias del decisor (o del experto) basado en el conjunto de ejemplos. Se trata por tanto de un modelo descriptivo, no prescriptivo, según se ha mencionado.

Una característica importante de la filosofía rough set aplicada a un problema de clasificación es que las inconsistencias mostradas en los

ejemplos de clasificación no se corrigen sino que las reglas producidas se clasifican en precisas o aproximadas. Dentro del contexto de clasificación, las reglas no determinísticas significan que bajo determinadas condiciones, no es posible asignar los objetos de manera unívoca a las categorías a menos que se solicite información adicional

Utilizando la terminología empleada de la inteligencia artificial o aprendizaje inductivo, el conjunto de ejemplos es una muestra de prueba (training) para aprender el comportamiento de las preferencias del experto (Grzymala-Busse, 1992).

Como consecuencia de lo expuesto anteriormente, tenemos que el conjunto de las reglas de decisión y la información sobre los atributos más significativos para la clasificación de los objetos puede considerarse como *una representación* del conocimiento adquirido por un *especialista* o *experto* sobre todos los casos/objetos contenidos en un sistema de información.

Hay que resaltar que en esta forma de representación no hay redundancias, tan típicas en las bases de datos reales.

Es más, los resultados obtenidos por el enfoque rough set se expresan de *forma similar al lenguaje natural humano*. Así, estas representaciones son muy fáciles de comprender por el usuario/analista y permiten justificar y explicar las conclusiones derivadas de análisis de los datos. Además, da la posibilidad para el analista de controlar el proceso de análisis de datos de manera simple. Tal posibilidad no es normalmente ofrecida por las técnicas tradicionales de análisis de datos, unido además a su baja carga computacional y posibilidad de interacción con un usuario en modo de diálogo.

El enfoque rough set, tal y como se puede comprobar en la bibliografía y en el capítulo 2 de este trabajo, ha dado buenos resultados en el análisis de sistemas de información pertenecientes a diferentes

dominios y sobre todo en problemas de clasificación con atributos múltiples. Las aplicaciones de este tipo de problemas corresponden a diversos ámbitos como son la medicina, farmacología, industria, finanzas, geología, ciencias sociales, etc. Algunos de éstas se comentan en la parte de aplicaciones en el capítulo 2.

b) Problemas de clasificación múltiple con atributos múltiples.

Este problema difiere del anterior en que los atributos de decisión son múltiples; por ejemplo, el conjunto de casos que hay que clasificar proviene de varios agentes. Para los mismos valores de los atributos de condición, las decisiones de clasificación pueden ser diferentes para algunos agentes y, por tanto, los modelos globales sobre las preferencias (políticas de decisión) pueden ser diferentes para cada uno de ellos.

Los ejemplos de clasificación se dan en forma de una tabla de decisión en la cual hay más de un atributo de decisión. Utilizando la metodología rough set para analizar una tabla de decisión, se pueden obtener los mismos resultados que para un problema de clasificación multiatributo pero relativo a cada agente en particular (atributos de decisión). Además, se puede medir el grado de consistencia de los agentes con la descripción de los objetos por medio del conjunto de atributos de condición, detectar y explicar las partes concordantes y discordantes de las políticas de decisión de los agentes, evaluar el grado de conflicto entre ellos, y construir los modelos de preferencias (reglas de clasificación) expresadas en términos comunes (atributos de condición) para facilitar la mutua comprensión de los agentes.

c) Descripción de objetos con atributos múltiples

Estos problemas están relacionados con la explicación de una situación de decisión. La metodología rough set se adapta especialmente bien cuando nuestro principal interés es buscar la mínima descripción posible en términos de atributos. Una descripción mínima permite un

minucioso análisis de los conflictos que es una importante cuestión a explicar. Veremos además que si los atributos son consecuencias de algunas decisiones, el análisis rough set permite descubrir las mínimas dependencias elementales entre las consecuencias; en algunas aplicaciones las consecuencias pueden ser interpretadas como relaciones "causa-efecto".

*c.1.) Descripción con atributos múltiples de una situación de decisión:*

El principal objetivo de una descripción con atributos múltiples de una situación de decisión puede formularse en el lenguaje de la filosofía rough set como la búsqueda, para cada descripción de objetos, de un sistema de información en términos del menor conjunto de valores de los atributos que diferencien exclusivamente a todos los objetos.

Como en el caso a) de clasificación con atributos múltiples, el método rough set ofrece varias ventajas como son la evaluación de la importancia de los atributos particulares, construcción del menor subconjunto de criterios independientes que aseguren la misma calidad de clasificación que el conjunto total, esto es, los reductos del conjunto de atributos y, si la intersección de los reductos no es un conjunto vacío, nos dará un núcleo de los criterios que no pueden eliminarse sin menoscabar la capacidad de aproximación de las reglas de decisión. Además, merece la pena mencionar que utilizando la metodología rough set se obtienen todas las posibles soluciones al problema considerado, esto es, todas las descripciones mínimas, cada una utilizando un conjunto diferente de atributos. Esto sugiere una optimización de la descripción, para el caso en que tengamos varias posibilidades de describir los objetos, y nos preguntemos cuál es la más útil siguiendo unos criterios preestablecidos.

La cuestión más importante que se puede abordar utilizando la teoría rough set es la búsqueda de lo que ocurre si algunos atributos

(valores de los atributos) no están disponibles, esto es, cómo se ve afectada la descripción de los objetos si faltan datos.

Por último, pero no por ello menos importante, puede ser de interés la descripción no de simples objetos, sino de conjuntos de ellos (subconjuntos del universo), en términos de valores de los atributos. Aquí resulta que una descripción exacta de un conjunto de objetos no es siempre posible y la descripción aproximada se presenta como un hecho. En este caso, se utilizan los conceptos de aproximación por encima y por debajo.

Por otro lado, otro problema importante que se puede abordar con la teoría rough set es el *análisis de conflictos*. Para ello es necesario asumir que en un conflicto al menos dos participantes entran en disputa por alguna cuestión. Dichos agentes pueden ser individuos, grupos, estados, partidos, etc. La relación de cada agente con el asunto en cuestión se presenta en forma de un sistema de información. Esto es, los objetos son los agentes que toman parte en el debate, los atributos son los asuntos o cuestiones que van a ser discutidas, y las entradas en la tabla son los valores de los atributos (opiniones, creencias, puntos de vista, votos, etc.) que son asignados únicamente a cada agente y a cada atributo, es decir, cada entrada correspondiente a la fila  $x$  y a la columna  $a$  representa la opinión del agente  $x$  sobre  $a$ . Clarificamos este último aspecto con un ejemplo:

Supongamos que cada atributo toma los valores 1, -1 y 0, indicando favorable, rechazo y neutralidad, respectivamente, hacia la cuestión que estamos considerando. Una vez definido esta clase de sistema de información se pueden definir tres relaciones binarias entre los agentes: *conflicto*, *neutralidad* y *alianza*. El agente  $x$  y el  $y$  están en alianza sobre la cuestión  $a$  si ambos son favorables o están en contra de  $a$ ; están en conflicto si uno está a favor de  $a$  y otro está en contra de  $a$ ; y están neutrales si al menos uno de ellos es neutral frente a  $a$ . La opinión de los agentes  $x$  e  $y$  puede extenderse de una cuestión hacia un conjunto

de cuestiones, mediante la definición de una función apropiada que asigne un “grado” de conflicto entre los agentes considerados en base a sus puntos de vista.

Formalizando estas ideas tenemos:

Primeramente vamos a definir rigurosamente las tres relaciones binarias básicas entre los agentes: *conflicto*, *neutralidad* y *alianza*. Para esta finalidad necesitamos la siguiente función auxiliar:

$$\Phi_a(x, y) = 1 \text{ si } f(a, x) \cdot f(a, y) = 1 \text{ ó } x = y,$$

$$\Phi_a(x, y) = 0 \text{ si } f(a, x) \cdot f(a, y) = 0 \text{ ó } x \neq y$$

$$\Phi_a(x, y) = -1 \text{ si } f(a, x) \cdot f(a, y) = -1$$

Esto significa que si  $\Phi_a(x, y) = 1$ , el agente  $x$  y el agente  $y$  tienen la misma opinión sobre la cuestión  $a$  (están aliados respecto a  $a$ ); si  $\Phi_a(x, y) = 0$ , al menos un agente  $x$  o  $y$  tiene una opinión neutral sobre la cuestión  $a$  (es neutral respecto a  $a$ ); y si  $\Phi_a(x, y) = -1$ , ambos agentes tienen opiniones diferentes sobre la cuestión  $a$  (están en conflicto respecto a  $a$ ). La función que hemos introducido arriba determina únivocamente las siguientes tres relaciones binarias:

$$R_a^+(x, y) \text{ si } \Phi_a(x, y) = 1$$

$$R_a^0(x, y) \text{ si } \Phi_a(x, y) = 0$$

$$R_a^-(x, y) \text{ si } \Phi_a(x, y) = -1$$

denominadas *relación de alianza*, *neutralidad* y *de conflicto* respectivamente.

Las relaciones  $R_a^+(x, y)$ ,  $R_a^0(x, y)$  y  $R_a^-(x, y)$  pueden extenderse a un subconjunto arbitrario  $B \subseteq Q$  de atributos mediante la definición de la función

$$\rho(x, y) = \sum_{a \in P} \Phi_a(x, y) / \text{card}(Q)$$



Obviamente,  $-1 \leq \rho(x,y) \leq 1$ . Si para el caso en particular que  $\rho(x,y) > 0$ , diremos que  $x$  e  $y$  están en *alianza (coalición)* sobre  $P$  en el grado  $\rho(x,y)$ , si  $\rho(x,y) < 0$ , diremos que  $x$  e  $y$  están en *conflicto* sobre  $P$  en el grado  $\rho(x,y)$  y si  $\rho(x,y) = 0$  diremos que  $x$  e  $y$  están *neutrales* respecto a  $P$ .

Las ideas señaladas se pueden extender y generalizar en muchas direcciones. Por ejemplo, es fácil definir las cuestiones muy (o muy poco) conflictivas y eliminarlas de un debate para resolver el conflicto. Sin embargo, hay que tener en cuenta que eliminando algunas cuestiones de un debate, pueden cambiar las relaciones entre las partes implicadas en el conflicto. Se puede también asignar a cada objeto su *fuerza*, y cada atributo un *peso*, obteniendo, de este modo, un modelo más interesante y realista de las situaciones conflictivas.

Esta clase de análisis puede ser útil de muchas formas, por ejemplo en negociaciones y resolución de conflictos, y contribuir de manera esencial a los procesos de toma de decisiones en presencia de conflictos. No vamos a entrar más en esta cuestión por quedar fuera de nuestro trabajo.

Finalmente, señalamos que en el caso de que los atributos se correspondiesen, por ejemplo, a árbitros que evaluaran a algunos competidores (objetos), entonces la cuestión sobre los conflictos no tendría sentido en relación a los objetos pero sí sería interesante en relación a los árbitros. Las consideraciones realizadas para aplicarse a este caso deberían invertirse.

El método rough set parece adaptarse muy bien a la descripción de situaciones de decisión, especialmente cuando nuestro principal interés es conseguir una descripción mínima en términos de atributos. Aquí es donde este enfoque muestra todo su potencial, y en contraste con otras teorías ofrece toda una gama de técnicas para investigar esta clase de situaciones.

c.2.) *Descubrir las dependencias entre las consecuencias de las decisiones.*

Este problema es similar al anterior, ya que en ambos casos los atributos no están divididos en atributos de condición y de decisión; sin embargo en el caso anterior nos preguntábamos cuestiones totalmente diferentes a las que nos vamos a cuestionar en este.

Aquí entran diversos tipos de situaciones de decisión pero sólo indicaremos dos de ellas para ilustrar los tipos de problemas que se podrían abarcar.

La primera es el caso en que los objetos se entienden como decisiones y los atributos como consecuencias (resultados, acciones, etc.) de estas decisiones. Por ejemplo la decisión de una familia de pasar sus vacaciones en diferentes países pueden implicar algunas consecuencias como es utilización de tarjeta o moneda local, requerimiento de algún equipamiento especial, etc. En este caso los países serían interpretados como objetos y las consecuencias como atributos.

Si los atributos se interpretan como consecuencias de algunas decisiones, no podemos quitar algunos de ellos, como en los problemas vistos anteriormente en las tres secciones anteriores, ya que todas las consecuencias son importantes y no pueden ser eliminadas de la situación de decisión. En su lugar, estaremos interesados en buscar las dependencias que existen entre las consecuencias, es decir, intentar descubrir cómo las consecuencias de nuestras decisiones están interrelacionadas unas con otras.

Otra situación sería que consideremos ahora un número finito de situaciones de decisión (estados) y en cada situación (estado) se tienen que emprender acciones específicas. Podemos imaginar el caso de una máquina o aparato que es controlado por una persona o por un controlador automático en la base de un estado corriente. El proceso de

control se presenta en forma de una tabla similar a las explicadas en la teoría. Los objetos de la tabla son estados de la decisión (por ejemplo, estados de la máquina) y los atributos representan acciones a realizar en cada estado específico. En este caso, cuando existe la dependencia funcional entre las acciones, algunas de ellas puede que tengan que ser desarrolladas cuando otras hayan terminado, mientras que acciones independientes pueden desarrollarse simultáneamente.

Se puede fácilmente comprobar con las dos situaciones expuestas previamente que en este caso nuestro principal interés se centra en las interdependencias de los atributos, en contraste con el apartado c.1) donde nos centrábamos en el descubrimiento de la mínima descripción de los objetos en términos de los valores de los atributos. Lo que representen los objetos y los atributos no es aquí lo importante y por tanto son posibles muchas interpretaciones de estos resultados en términos de situaciones de decisión.

## **Capítulo 2**

### **APLICACIONES DEL MÉTODO ROUGH SET**

## I.- INTRODUCCIÓN

La predicción del fracaso empresarial es un campo científico que cuenta con muchos académicos y profesionales que han venido trabajando por lo menos en las últimas tres décadas. Las organizaciones financieras como bancos, instituciones de crédito, clientes, ... están también interesadas en la predicción del fracaso de las compañías.

Los intentos para la predicción de las crisis empresariales comienzan con el uso de métodos empíricos propuestos por los grandes bancos de EEUU, como el método de “las cinco C” (carácter, capacidad, capital, condiciones y cobertura), el método “LAPP” (Liquidez, Actividad, rentabilidad (Profit) y Potencial) y el método “credit-men” (Zopounidis, 1995).

Luego, se desarrollaron los métodos basados en los ratios financieros para el problema que estamos considerando. Los ratios han sido considerados como indicadores objetivos del fracaso de una empresa (Beaver, 1966, Courtis, 1978 y Altman, 1993).

La investigación de los ratios financieros también denominada enfoque estadístico univariante, dio lugar a los métodos para la predicción de las quiebras en las compañías basados en el análisis estadístico multivariante. Ya en 1968, Altman propuso la utilización del análisis discriminante para la predicción del riesgo de crisis (obtuvo una función discriminante con cinco ratios financieros).

En consecuencia, el uso de este método ha continuado extendiéndose hasta el punto de que hoy en día podemos hablar de modelos discriminantes para la predicción del fracaso empresarial. Al mismo tiempo que esta generalización del uso del discriminante han aparecido estudios que critican su validez.

Desde el trabajo de Altman (1968), han surgido numerosos trabajos que proponen otros métodos que superen las desventajas del análisis discriminante y que den una mejor capacidad de predicción.

Recientemente, se han desarrollado nuevos métodos en la predicción de las crisis que se apoyan en los avances en las ciencias de la computación y la información. Estos ofrecen a los directores de las instituciones y empresas financieras ayuda importante en la selección y evaluación de empresas viables para la financiación.

Muchos métodos de los desarrollados se basan en la metodología de ayuda a la decisión con criterios múltiples. Estos métodos clasifican a las empresas en categorías de acuerdo al riesgo de fracaso empresarial. El uso de esta metodología en la predicción del riesgo de quiebra supera muchos de los problemas que existen cuando se utiliza el análisis discriminante.

Finalmente han ido apareciendo herramientas basadas en la inteligencia artificial, que son las que en un principio se denominaron sistemas expertos, posteriormente sistemas de apoyo al experto o sistemas de apoyo a la decisión basados en el conocimiento. Estos sistemas se construyeron para la valoración de una firma, para la evaluación de la concesión de un préstamo, para diagnóstico financiero de una compañía y para analizar proyectos de creación corporativa.

En este capítulo se va a aplicar el enfoque rough set explicado en las secciones anteriores, al problema del análisis y predicción de las crisis empresariales. Según hemos visto esta teoría desde su introducción por Pawlak en 1982 ha atraído a muchos investigadores de todo el mundo que han contribuido a su desarrollo durante la última década. Esta herramienta ha demostrado ser de gran utilidad en el análisis de sistemas de información (tablas de información financiera) que describen a un conjunto de objetos (empresas) mediante un conjunto de atributos valorados (ratios financieros o variables cualitativas).

Tal y como comprobaremos, en el caso de que el conjunto de objetos este sujeto a la opinión de un experto, el enfoque rough set nos permite:

- reducir el conjunto de atributos (cuantitativos o cualitativos) a un subconjunto que nos asegure la misma aproximación a la clasificación del experto que el conjunto total de atributos.
- derivar reglas de decisión del sistema de información con vistas a explicar la política de toma de decisiones del experto.

Estos resultados, tanto la reducción de atributos como la obtención de un algoritmo de decisión, tienen una gran importancia desde el punto de vista práctico. Por ejemplo, en medicina, la reducción puede llevar a la eliminación de pruebas clínicas que no sean necesarias y que sean dolorosas para los pacientes.

Utilizando la terminología de los analistas de toma de decisiones con atributos múltiples (Roy, 1985), la evaluación del riesgo de quiebra es un problema de clasificación multiatributo que se reduce a una cuestión: como asignar a cada empresa a una categoría de riesgo de entre un conjunto de ellas previamente definidas, dada su descripción multiatributo.

Los estudios que vamos a comentar en la sección siguiente muestran la capacidad de la teoría rough set en la predicción del fracaso empresarial, incluso en comparación con otros métodos estadísticos (discriminante y logit).

La utilidad práctica del concepto de rough set ha sido probada en muchas aplicaciones demostrándose su provecho como herramienta para el análisis de amplias categorías de problemas de decisión con atributos múltiples.

En esta introducción, y a modo indicativo, solo vamos a señalar algunas de las aplicaciones prácticas para problemas distintos del fracaso

empresarial y verificar la utilidad de esta herramienta en los problemas de decisión.

En la sección 2.2. comentaremos tres trabajos referidos al problema de la predicción de las crisis empresariales.

Hay que tener en cuenta que el primer software que se desarrolló para implantar en un ordenador la teoría rough set fue el RoughDas y el RoughClass desarrollados en la Universidad de Tecnología de Poznan (Polonia) y son los antecesores directos del ROSE (que es el software que hemos utilizado en la parte empírica del trabajo) por lo que muchas de los trabajos que vamos a comentar están realizados con estos programas. Una descripción más amplia de los mismos la hemos desarrollado en el capítulo 3.

Uno de los campos donde más se ha aplicado la teoría Rough Set, y con más éxito, es el ámbito de la medicina. En esta área los resultados encuentran una gran acogida entre los especialistas; en concreto, los modelos de pacientes que se crean y los algoritmos de decisión de ellos derivados son apreciados como herramientas útiles para ayudar en la asignación de terapias.

Además la reducción de atributos no significativos puede llevar a la eliminación de pruebas clínicas que sean dolorosas para los pacientes y en los diagnósticos puede reducir el coste y el tiempo de los mismos.

Dentro del campo de la medicina existe un estudio relativo al análisis de un sistema de información con datos sobre pacientes que sufrían de úlceras duodenales tratadas con vagotomía altamente selectiva (HSV) (Fibak et al., 1986). El *sistema de información HSV* se compuso de 122 pacientes tratados con HSV y descritos a través de 11 atributos (tanto cualitativos, como sexo o complicación de la úlcera, como cuantitativos, los relativos a la anamnesis o los resultados del laboratorio) preoperatorios. El conjunto de pacientes se clasificó en cuatro clases de



acuerdo a los resultados a largo plazo de la operación (las clases eran: excelentes resultados, buenos resultados, resultados satisfactorios y resultados no satisfactorios). Aplicando la metodología rough set la descripción se redujo a cinco atributos relevantes que mantenían una calidad aceptable de clasificación. Los atributos reducidos se basaban en tests o pruebas que podían dar efectos secundarios no deseados para los pacientes. La aplicación de las 44 reglas de decisión obtenidas de las aproximaciones por debajo de las clases de buenos y malos resultados de la operación a 70 pacientes nuevos dio un incremento de buenos resultados de la HSV del 82% al 93%.

Otra aplicación se refiere al análisis de las relaciones entre la estructura química y la actividad antimicrobial de 201 compuestos cuaternarios imida. Los compuestos se describieron mediante ocho atributos relacionados con la estructura y se dividieron en cinco clases de actividad. El reducto de los atributos que se descubrió utilizando el sistema RoughDAS se compuso de cuatro atributos. Se obtuvieron dos conjuntos de reglas de decisión, que incluían 22 y 35 reglas, respectivamente, y que daban claras recomendaciones (en términos de las características relevantes de la estructura) sobre cómo diseñar nuevos compuestos activos antimicrobiales.

Dentro del campo del diagnóstico técnico, la teoría rough set se ha aplicado al análisis de la capacidad de diagnóstico de síntomas vibroacústicos. Este enfoque se ha convertido en una buena herramienta para la comparación objetiva de diferentes métodos de definir los valores límites de los síntomas tanto para atributos de ruido como para atributos de vibración en la diagnosis de comportamientos ondulatorios. Como resultado del estudio se halló una clara superioridad de los síntomas de vibración sobre los de ruido y se construyó una clasificación del estado técnico del comportamiento ondulante que consistió en 14 reglas de decisión que utilizaban tres (sobre doce) síntomas relevantes.

Otro estudio interesante es el relativo al uso del método rough set para la reducción de datos en una red neuronal que clasifica fotografías microscópicas de tumores cerebrales. La aceleración del tiempo de aprendizaje debida a la reducción está por encima de 4.72 veces. Un estudio colateral de este último concluye que el número mínimo de neuronas en el estrato escondido es igual al número cardinal del reducto más pequeño. Estos resultados son prometedores y demuestran la utilidad de la metodología rough set como herramienta para reprocesar datos de redes neuronales.

Estos son solo algunas de las aplicaciones de la teoría y muchas más están recogidas en Slowinski (ed.), (1992).

Antes de entrar en el ámbito financiero, y ya dentro de las ciencias sociales hemos de comentar que Nurmi et al., (1996) han aplicado los conceptos de la teoría rough set al problema de la elección colectiva ya que esta teoría “proporciona un modo sistemático de descubrir las propiedades cruciales y diferenciadoras de los distintos sistemas de voto a través de los conceptos de reducto y de núcleo” (Nurmi et al., 1996, p. 265).

En el campo financiero a parte de la aplicación a la predicción del fracaso empresarial, existen algunos intentos de predecir los comportamientos del mercado bursátil. Uno de estos intentos se basa en una variante de la metodología rough set que es el denominado *modelo rough set de precisión variable* (Ziarko, 1993), modelo que incorpora el ROSE. Este modelo se ha aplicado para analizar los datos mensuales recogidos de la bolsa durante un período de diez años. El objetivo de dicho análisis era identificar las relaciones dominantes entre las fluctuaciones de los indicadores de la bolsa y los precios bursátiles.

El origen de este objetivo vino motivado porque en la década de los ochenta surgieron numerosos intentos desde la esfera de la inteligencia artificial de intentar, de alguna manera, predecir las fluctuaciones en el

mercado bursátil o buscar algún posible indicio en ellas que nos pudiese facilitar dicha predicción. La tarea no era fácil y se hacía necesario una gran interacción entre los sistemas de análisis de datos incorporados en los ordenadores, los datos para la construcción de un modelo y los expertos en bolsa (brokers).

El objetivo de estos sistemas expertos era dar a los brokers de una determinada firma alguna ventaja sobre sus competidores que se tradujese en beneficios extras.

Los primeros intentos para realizar dicho objetivo desde el ámbito de la inteligencia artificial se hicieron a través de redes neuronales pero había una gran dificultad en interpretar los resultados que se obtenían. De aquí surge la necesidad de desarrollar metodologías y herramientas que aumentasen la comprensión del funcionamiento del mercado y diesen predicciones más precisas.

Los métodos estadísticos fueron descartados porque su capacidad predictiva se veía muy mermada por la interacción entre variables que no seguían distribuciones de probabilidad conocidas como es el caso de los datos y variables que intervienen en el proceso de análisis y predicción del mercado de valores.

Por lo tanto el problema consistía en aplicar técnicas de descubrimiento del conocimiento para identificar reglas predictivas fuertes de los datos económicos y bursátiles que fuesen una auténtica orientación de lo que ocurría durante un determinado periodo de tiempo en la bolsa.

Para este propósito de comparación, se obtuvieron reglas tanto precisas como imprecisas, fuertes y débiles que después fueron evaluadas por un broker.

Por reglas fuertes entendemos, tal y como se ha mencionado anteriormente, reglas que reflejen patrones altamente repetitivos que ocurren en los datos. No tienen por qué ser necesariamente precisas o

determinísticas, ya que pueden estar asociadas con fracciones de probabilidad de los resultados predichos. Lo que si deben ser son correctas o casi correctas en el sentido de que reflejen relaciones reales existentes en el sistema económico.

Para medir la fuerza de las reglas se puede hacer en términos de objetos (datos registrados) que satisfacen la parte condicional de la regla. Las reglas fuertes representan normalmente patrones que siguen los datos y probablemente son regularidades generales y ciertas que siguen los mismos. Además en teoría, si la muestra es representativa y aleatoria de todas las combinaciones posibles de los indicadores del mercado se podría decir, utilizando la teoría de la probabilidad, que cuanto más fuerte sea una regla más probable es que sea cierta o se aproxime a un hecho cierto.

Con la valoración efectuado por el experto (broker) se pretendía depurar del “ruido” presente en las reglas, esto es, aquellas reglas que son claramente una contradicción con la experiencia del experto y aquellas de cuya fuerza relativa se sigue de que los datos disponibles no son adecuados.

En el problema del mercado de valores, el universo es el conjunto de todos los posibles estados del mercado que se representan en diferentes instantes del tiempo a través de las mediciones de los indicadores del mercado. En la aplicación en concreto, los datos utilizados eran representativos del mercado bursátil canadiense para el periodo 1980-1990 sobre una base mensual. Dichos datos pertenecían a tres categorías: datos bursátiles, indicadores económicos y datos individuales de las compañías cotizadas seleccionadas. Por tanto la capacidad de distinguir diferentes estados en el mercado queda restringida a los indicadores del mercado disponibles. En consecuencia, dos estados con los mismos indicadores no son diferenciables lo cual conduce a la división de los estados en clases de estados idénticos (en el sentido de no ser

distinguibles en base a los indicadores del mercado considerados) o, lo que es lo mismo, en conjuntos elementales.

Un total de 120 vectores con 40 valores de los atributos se recogieron y utilizaron para el análisis. Los datos originales fueron discretizados mediante símbolos que representaban el porcentaje de variación de un valor en el período actual sobre el valor registrado en el mes previo. Dichos intervalos utilizados para discretizar fueron establecidos por el broker y se llevaron a un fichero ASCII y de ahí al sistema DATALOGIC/R donde estaba incorporado la variante del rough set que se utilizó. Antes de utilizar dicho modelo hubo que definir cada variable, darle un nombre, a que tipo pertenece y su longitud.

También y siguiendo con el modelo hubo que definir el parámetro de “roughness” (aproximación) o el parámetro BETA de probabilidad mínima de aceptación de una regla.

Las reglas obtenidas se lograron variando el parámetro “roughness” para así ver la información desde diferentes puntos de vista y comprobar como variaba. Colocando dicho parámetro en un nivel bajo, se generan reglas más generales y fuertes. Situando el parámetro en un nivel alto (1), se generan reglas exactas (es decir con una probabilidad estimada igual a uno) pero relativamente débiles. Por tanto en todas las reglas que se generaron el parámetro BETA se situó al menos en 0.55.

Del trabajo realizado se obtuvieron las siguientes conclusiones:

No todas las reglas obtenidas eran de la misma calidad ya que algunas eran difíciles de interpretar.

La valoración de dichas reglas demostró que las reglas fuertes (aquellas que eran verificadas por muchos casos) en esencia confirmaban la experiencia del experto mientras que las débiles eran con frecuencia difíciles de interpretar. Esto sugirió que la utilización de la fuerza de la regla fuese el primer criterio para la selección de reglas predictivas que

fuesen útiles y apuntaba a una posible línea de investigación sobre cuál debía ser la fuerza que debía poseer una regla para considerarla correcta.

Otra conclusión es que para mejorar el proceso de selección de reglas, debía mejorarse en modo significativo los datos ya que no solo se demostraron ser insuficientes sino que su calidad también debía mejorarse. Además se verificó la necesidad de incorporar otros conceptos como las ganancias por dividendos sobre una base mensual.

También se vio la posibilidad de utilizar los tres grupos de datos pero para generar reglas que explicasen qué había ocurrido en el mes anterior para que en el actual la bolsa, subiese bajase o permaneciese igual. De esta manera y considerando muchos meses pasados se podrían generar reglas temporales lo que supondría una capacidad de predicción de la tendencia intermedia. Las predicciones a estos niveles intermedios podrían dar una gran ventaja a los inversores.

Sin embargo, y como punto pendiente de dicho trabajo, no se efectuó la validación de las reglas mediante la utilización de estas sobre un conjunto independiente de datos porque para ello se necesitaban grandes cantidades de datos que no estaban disponibles.

## **II.- LA PREDICCIÓN DE LAS CRISIS EMPRESARIALES**

En esta parte del trabajo pretendemos comentar cuatro aplicaciones de la teoría rough set en relación a la predicción del fracaso empresarial.

La investigación para la predicción del fracaso empresarial a través de la metodología Rough Set es un acontecimiento reciente. Una búsqueda entre los artículos publicados en revistas en los últimos años revela la existencia de cuatro únicos estudios empíricos sobre la predicción del fracaso empresarial que son los que pasamos a resumir.

El primero de ellos es el trabajo de Slowinski y Zopounidis (1995). Aplicaron este enfoque a un problema real: un banco griego estaba interesado en invertir sus fondos en las mejores firmas.

El primer elemento en la valoración de una firma es la evaluación de su riesgo. Se eligió una muestra de 39 compañías que quebraron entre los años 1986-1990 en Grecia. La selección se hizo en trece sectores y teniendo como criterios el que hubiese disponibilidad de datos y que hubiesen estado funcionando durante más de cinco años. Se recopilaron los estados financieros de cinco periodos previos a la quiebra (años - 5 a -1, siendo el año -1 el año previo a la quiebra) Por tanto el año de quiebra, que podríamos denominar año 0, no coincide obviamente para todas las empresas que quebraron.

Las 39 empresas fracasadas se emparejaron con otras 39 que no fueron a la bancarrota (empresas que denominaremos sanas). Estas últimas fueron elegidas teniendo en cuenta el que pertenecieran al mismo sector, que tuvieran un activo similar y mismo número de empleados para el año previo a la quiebra. Esta forma de seleccionar la muestra ha sido utilizada por numerosos investigadores (Altman (1968), Beaver (1966), entre otros).

Se hizo una partición triconómica de las mismas. La clasificación se representó en el sistema de información mediante el atributo de decisión

d. Así,  $d=1$  se estableció para las firmas “aceptables” (bajo nivel de riesgo);  $d =2$  para las firmas “inciertas” para las que la información disponible no es suficiente y requieren un estudio adicional y  $d=3$  para las firmas “inaceptables” (grupo de las quebradas).

Las firmas fueron evaluadas de acuerdo a un total de 12 atributos, 6 de ellos cuantitativos (ratios financieros) y los otros seis cualitativos.

Los ratios financieros eran:

$A_1 = \text{Beneficios antes de intereses e impuestos} / \text{Activo total}$

$A_2 = \text{Ingresos netos} / \text{Activo neto}$

$A_3 = \text{Pasivo total} / \text{Activo total}$

$A_4 = \text{Pasivo total} / \text{Cash Flow}$

$A_5 = \text{Gastos de Intereses} / \text{Ventas}$

$A_6 = \text{Gastos Generales y Administrativos} / \text{Ventas}$

La selección de los ratios financieros se basó en dos características principales:

a) su utilidad en estudios predictivos previos (Altman, 1968 y Courtis, 1978) y

b) la experiencia en decisiones financieras pasadas, el conocimiento y las preferencias del director financiero.

Los atributos cualitativos fueron:

$A_7 = \text{Experiencia Laboral de los directivos}$

$A_8 = \text{Nicho de mercado (posición)}$

$A_9 = \text{Estructura Técnica (medios)}$



$A_{10}$ = Organización (personal)

$A_{11}$  =Ventajas competitivas especiales de las empresas y

$A_{12}$ = Flexibilidad de mercado.

Los atributos 7 a12 revelan la necesidad de la dirección del banco de evaluar el riesgo de bancarrota teniendo en cuenta alguna información general cualitativa de las firmas además de los datos financieros. Estos atributos daban un análisis cualitativo del fracaso. Los seis atributos cualitativos se introdujeron en el modelo de acuerdo a una escala ordinal de 4 (el mejor) a 1 (el peor).

Para la aplicación de la investigación rough sets se formó un sistema de información para representar la experiencia del banco que consistió en transformar los valores continuos de los atributos cuantitativos en términos cualitativos. De esta manera todos los atributos quedaban codificados utilizando números naturales.

Esta transformación de atributos cuantitativos en atributos cualitativos se hizo de acuerdo a *normas* que seguían la experiencia del director financiero y los estándares del análisis financiero corporativo.

Esta transformación implica una división del dominio original en subintervalos y una asignación de términos nominales a ellos. Los valores frontera de los subintervalos es lo que denominamos *normas*. Surgen normalmente de convencionalismos, hábitos o de asignación subjetiva.

La utilización de *normas* que transformen los atributos cuantitativos en términos cualitativos no se impone por el rough set pero sí para una interpretación práctica de los atributos cuantitativos. Incluso si un atributo representa una medida continua, como es un ratio financiero, o la presión sanguínea en medicina, el experto normalmente interpreta los valores de los atributos en términos cualitativos, esto es, bajo, medio o alto. Además, y dado que las normas se utilizan desde el principio del problema y hasta

el final en la explicación de las reglas de decisión, no falsean la imagen original de la situación de decisión.

Todos los cálculos del estudio se transformaron utilizando el sistema "RoughDAS" desarrollado por Slowinski y Stefanowski (1992).

Las aproximaciones se hicieron para cada categoría particular de empresas por el conjunto Q de todos los atributos y se calculó la precisión de cada aproximación. La precisión de las aproximaciones fue perfecta, esto es, igual a uno y la calidad de la clasificación fue también igual a uno. El resultado es muy significativo para el director financiero, porque a pesar de un gran sistema de información las empresas están muy bien discriminadas entre ellas.

El siguiente paso del análisis "rough set" fue la construcción del mínimo subconjunto de atributos independientes que asegurasen la misma calidad de clasificación que la totalidad del conjunto Q, esto es los reductos de Q en el sistema de información codificado.

La intersección de los reductos nos da el núcleo. En este caso se compuso solamente del atributo  $A_7$  lo que significa que es el atributo más significativo y que no puede eliminarse de Q sin disminuir la calidad de la aproximación de la partición. Cualquier otro atributo aparece al menos en un reducto pero no es indispensable para la aproximación. El número cardinal de los reductos era de cuatro o cinco, así que en cualquier caso ocho o siete atributos eran redundantes y podían ser eliminados sin ninguna consecuencia en la aproximación.

Para seleccionar uno de los reductos se eligió el núcleo y el par que daba la más alta calidad de clasificación. Elegido el par se eligió la terna que daba la más alta calidad de clasificación, y así sucesivamente, hasta que la calidad era igual a uno. El subconjunto de atributos que simultáneamente alcanzaba la calidad igual a uno fue elegido como el mejor reducto.

Con el reducto se pasó de un sistema inicial de información codificado de 12 a 4 columnas. El sistema de información reducido puede verse como una tabla de decisión. Dado que el reducto daba una calidad de clasificación igual a uno, todas las reglas son determinísticas, esto es, describen únicamente la decisión que ha de tomarse cuando se satisfacen algunas condiciones.

Las quince reglas obtenidas constituyen un algoritmo de clasificación siendo una representación no redundante de la experiencia del banco y reunidas en un sistema de información codificada. En este algoritmo solo se utilizaron 31 descriptores lo que supone menos del 7% que aparecían en el sistema de información inicial (468).

Utilizando el software "RoughDAS" se obtuvo (ya que incluye esa posibilidad) el algoritmo de clasificación más compacto del estudio.

Se caracterizaba por el mínimo número de descriptores pero el número de atributos utilizados por este algoritmo no se limita a los atributos de un reducto dado. Este algoritmo contenía 11 reglas y utilizó 9 atributos pero solamente 23 descriptores, esto es menos del 5% de los descriptores originales. El algoritmo de 11 reglas es una combinación diferente de los valores de los atributos que pueden clasificar en tres grupo definidos *a priori*.

Las reglas de los dos algoritmos eran diferentes (uno da quince reglas y el otro once) aunque algunas eran comunes a ambos.

La muestra de 39 firmas utilizada en el estudio para derivar los algoritmos de decisión puede considerarse como una muestra de prueba utilizada para revelar la política financiera del banco que aplicada en la selección de las empresas viables. Los dos algoritmos de clasificación podrían ser utilizados para evaluar cualquier otra empresa que busque financiación de dicho banco. La precisión predictiva debería contestarse en el curso de la verificación práctica ya que los algoritmos de

clasificación fueron desarrollados en primer lugar para explicar la experiencia previa del banco.

El segundo estudio es el de Dimitras *et al.*, (1999).

En cuanto a los datos de la muestra son los del estudio anterior. Pero además se seleccionó una segunda muestra o muestra de prueba (“testing sample”) utilizando la misma filosofía. Esta segunda muestra constaba de 19 empresas fracasadas del periodo 91-93 y de otras 19 no fracasadas.

De los estados financieros se calcularon 28 ratios. No se consideraron características cualitativas debido a la imposibilidad de conseguirlas.

Para la aplicación de la investigación rough set, el decisor (director financiero de un banco griego) jugó un papel, muy importante en:

- la elección de los ratios financieros (atributos) que entrarían en la tabla de información,
- la transformación en forma discreta de los atributos continuos, mediante el establecimiento de normas que dividieran el dominio original de los atributos en subintervalos,
- la selección del reducto satisfactorio de atributos de entre todos los reductos calculados de la muestra, y
- la validación del conjunto de reglas de decisión sobre la muestra de prueba.

El decisor seleccionó doce ratios (de entre los 28 calculados) en un principio para que entrasen en la tabla de decisión. La elección fue justificada por:

- a) Los ratios seleccionados representaban las tres categorías propuestas por Courtis (1978), ratios de rentabilidad, gestión y solvencia.
- b) Un análisis preliminar de las características de los dos grupos de empresas (sanas y fracasadas) efectuado por el decisor.
- c) Las preferencias del decisor, su conocimiento, su experiencia...

Los ratios seleccionados fueron los siguientes:

$$A_1 = \text{Ingresos netos} / \text{Beneficio Bruto}$$

$$A_2 = \text{Beneficio Bruto} / \text{Activo Total}$$

$$A_3 = \text{Ingresos netos} / \text{Activo total}$$

$$A_4 = \text{Ingresos netos} / \text{Activo neto}$$

$$A_5 = \text{Activo circulante} / \text{Pasivo circulante}$$

$$A_6 = \text{Activo realizable} / \text{Pasivo Circulante}$$

$$A_7 = (\text{Deuda a largo plazo} + \text{pasivo circulante}) / \text{Activo Total}$$

$$A_8 = \text{Activo neto} / (\text{Activo neto} + \text{deuda a largo plazo})$$

$$A_9 = \text{Activo neto} / \text{Activo fijo neto}$$

$$A_{10} = \text{Existencias} / \text{Capital circulante}$$

$$A_{11} = \text{Pasivo circulante} / \text{Total activo}$$

$$A_{12} = \text{Capital circulante} / \text{Activo neto}$$

El decisor discretizó los ratios financieros, variables de tipo continuo, de acuerdo a su conocimiento y a un análisis preliminar de las empresas. Este proceso se hace necesario debido a que la precisión de los ratios financieros es bastante dudosa, y dificulta el extraer conclusiones

generales de los datos en términos de dependencias, reductos o reglas de decisión.

Las empresas para las que los valores de estos ratios financieros estaban en los mismos subintervalos se suponía que tenían el mismo comportamiento y características. Los códigos que se utilizan para referirse a un subintervalo no representan ningún orden de preferencias. La teoría rough set no tiene en cuenta el orden, ya que trabaja con la relación de no diferenciación, es más la selección de los códigos no afecta a los resultados obtenidos.

Utilizando estos subintervalos y el principio “cuánto más alto sea el código, mejor subintervalo”, se obtuvo una tabla de información codificada. Dado el principio que se aplicó, para los atributos decrecientes (aquellos en que los valores más bajos son mejores) se les dio un código en orden inverso. Fue además una oportunidad para hacer correcciones en la escala en el caso en que las preferencias del decisor no estuvieran en concordancia con la secuencia de aumento o disminución de los subintervalos.

La tabla codificada, que es la que se utilizó para un posterior análisis, constaba de las 80 empresas descritas por los 12 atributos codificados (ratios financieros), utilizando los datos del año previo a la quiebra (año  $-1$ ) y asignadas de forma binaria a una clase de decisión (sana o fracasada, representado por 1 ó 0 respectivamente).

El análisis rough set de la información codificada de la tabla se desarrolló utilizando los sistemas de RoughDAS (Slowinski y Stefanowski, 1992) y Pro FIT (Mienko et al., 1996). La totalidad del conjunto de atributos dio una aproximación perfecta de las clases de decisión además de la calidad de la clasificación (ambas igual a uno). El núcleo de atributos estaba vacío. Esto indica que ningún único atributo es absolutamente necesario para la aproximación de las dos clases de decisión. Si el núcleo no hubiese sido vacío indicaría que existen atributos en el sistema que

son indispensables desde el punto de vista discriminante, porque si no se incluyen cualquiera de los atributos contenidos en el núcleo conduce inmediatamente a una disminución de la calidad de la aproximación. Por otra parte, un núcleo no vacío ayuda a determinar los atributos más importantes además de las aproximaciones de las clases con las que está relacionado.

Se obtuvieron 54 reductos de la tabla de información codificada cada uno de los cuales contenía entre 5-7 atributos, lo que respecto a los 12 originales supone un reducción importante. Este resultado nos da una idea de la ayuda que supone la aplicación de este método ya que los reductos contienen unos pocos atributos pero asegurando la misma calidad de la aproximación que el conjunto total de atributos  $Q$ . El atributo  $A_{11}$  es el que con más frecuencia aparece (47) y el que aparece con menos frecuencia es el  $A_3$  (15).

Se presentaron los reductos al decisor al que se le pidió que seleccionara el que mejor se ajustara a sus preferencias. La selección se hizo teniendo en cuenta dos criterios:

- a) El reducto debía contener el menor número de atributos como fuese posible, y
- b) Debía contener los atributos considerados más significativos para la evaluación de las empresas.

Elegido el reducto, el resto de atributos fueron eliminados de la tabla de información codificada y se derivaron un conjunto de reglas de decisión tanto de la tabla de información codificada como de la tabla reducida.

Las reglas de decisión derivadas de la tabla de información codificada son reglas exactas, dado que la calidad de la clasificación es igual a uno y, en consecuencia, las fronteras de las clases de decisión estaban vacías.

Las reglas de decisión se derivaron de la tabla de información codificada y reducida según estas tres posibles estrategias:

- A) El mínimo conjunto de reglas que cubren todos los objetos de la tabla de decisión se compuso de 13 reglas.
- B) El conjunto exhaustivo de reglas consistente en todas las posibles reglas de la tabla de decisión y compuesto de 45 reglas se descartó por la poca disponibilidad y la pobre precisión de la clasificación en el test de reclasificación.
- C) El conjunto de reglas “fuertes” y discriminantes que son las que estaban soportadas por al menos 8 objetos cada una y se obtuvieron uno total de 10 reglas.

En general, cuanto más corta es una regla más general es. En los tres conjuntos de reglas de decisión, el número de reglas de decisión que describían a las empresas quebradas era mayor que el número de reglas que describían a las sanas. Esto significa que cuesta más generalizar la descripción de las empresas quebradas que las de las sanas.

Para probar las reglas de decisión, el decisor suministró la información necesaria para la Relación de Proximidad Valorada, es decir los niveles y pesos correspondientes a los cinco atributos del reducto.

Esta información refleja directamente la opinión del decisor sobre la importancia de los atributos para el desarrollo de una empresa.

Los tres conjuntos de reglas de decisión fueron probados primero en las empresas de la muestra de aprendizaje, y aplicados a los datos de los años -2, -3, -4, -5, es decir, 2, 3, 4 y 5 años antes del estado descrito en la tabla de información.

Luego, los tres conjuntos de reglas de decisión se utilizaron para la clasificación de las empresas de la muestra de prueba y para los años -1, -2 y -3.



Todos los test de reclasificación fueron probados dos veces: la primera sin y luego con la técnica de Relación de Proximidad Valorada para la clasificación de objetos que no se ceñían a ninguna regla durante la primera prueba. Esto mostraba una clara percepción de la eficiencia de esta técnica, ya que por término medio, un 60% de los objetos no clasificados por no ceñirse exactamente a ninguna de las reglas fueron correctamente clasificados por esta técnica a las reglas más cercanas, el 40% fue clasificado incorrectamente. Esto es mejor resultado que una clasificación al azar de los objetos.

Otra conclusión general es que existían dos reglas débiles (solo soportadas por un objeto) y no fueron utilizadas en ningún test de reclasificación. Esto significa que dichos objetos que las soportaban eran marginales.

Los resultados muestran que yendo hacia atrás del año  $-1$ , la precisión de la clasificación va normalmente disminuyendo de manera más rápida para las firmas quebradas que para las sanas. Esto era esperado porque cinco años antes del año de la quiebra, los problemas típicos de las empresas quebradas podrían no haberse presentado todavía, mientras que las empresas sanas permanecían estables y con un buen desarrollo durante los cinco años. Además la precisión de la clasificación para la muestra de prueba no fue tan buena como para la muestra de aprendizaje.

En general los resultados obtenidos son bastante satisfactorios ya que para la muestra de las 39 empresas y para todos los conjuntos de reglas se obtuvieron, por término medio, un porcentaje de acierto en la clasificación de más del 90% para el año  $-1$  y más del 75% para el año  $-5$ . En la muestra de prueba los resultados empeoran pero son aceptables ya que haciendo las mismas consideraciones que en el caso anterior los resultados de acierto van del 75% para el año  $-1$  y del 50% para el año  $-3$ . Como nota indicativa mencionamos, que mientras el conjunto de reglas “fuertes” y en parte discriminantes va mejor en la muestra de prueba, el

conjunto de reglas mínimo va ligeramente mejor en la muestra primera. Pero las diferencias son muy pequeñas.

Otra parte del estudio consistió en la comparación de los resultados de este método con el análisis discriminante.

Aunque la filosofía de ambos es muy diferente, su comparación sobre un conjunto de datos comunes está bien fundamentada porque las dos metodologías pueden aplicarse al problema de la predicción del fracaso empresarial.

La elección del análisis discriminante se efectuó debido a que este método es el primero y el más ampliamente utilizado para la clasificación y predicción en los problemas de gestión financiera (Altman et al., 1981; Altman, 1993; Dimitras et al., 1995).

El análisis discriminante es un método paramétrico y tiene que respetar varias hipótesis de tipo estadístico como son la normalidad de la distribución de las variables, la igualdad de las matrices de varianzas y covarianzas de las variables para las dos clases de empresas y especificación de las probabilidades a priori y/o los costes de error en la clasificación. Es muy costoso encontrar todas estas hipótesis en la práctica por lo que su validez ha sido cuestionada por numerosos autores.

A diferencia del análisis discriminante, la investigación rough set no necesita ninguna suposición sobre los datos previa al análisis (Krusinska et al., 1992 y Stefanowski, 1992 hacen una discusión crítica de los dos métodos).

Para comparar estos dos métodos, se usaron los datos originales y no los que estaban en forma discreta. Se construyeron dos funciones discriminantes, una de la tabla de información completa (con los 12 atributos) y otra con los atributos reducidos (los del reducto).

De esta comparación se observaron mejores resultados de la función discriminante derivada de la tabla de información completa. Sin embargo estos resultados son menos satisfactorios que los obtenidos utilizando la investigación rough set. Es más, desde el punto de vista del usuario, las reglas de decisión expresan las dependencias entre las características financieras de una empresa y su expresión es mucho mas adecuada que una función discriminante. Las reglas de decisión “hablan” el lenguaje natural de los ejemplos de decisión que dio el experto, mientras que la función discriminante da únicamente una visión global.

Los resultados dan, en definitiva, una evidencia de la capacidad del enfoque rough set para responder de manera eficiente al problema de la predicción del fracaso empresarial, convirtiéndose en una alternativa fidedigna al análisis discriminante.

Posteriormente, en el mismo estudio se hizo una comparación del enfoque rough set con el análisis logit.

Muchos investigadores financieros han utilizado el análisis logit para superar los problemas de aplicación del análisis discriminante a los problemas de clasificación del mundo real. La principal ventaja del análisis logit sobre el análisis discriminante es que no requiere que las variables independientes sigan una normal multivariante. Más aún, da mucha más información útil al analista financiero en la predicción de la quiebra, ya que no solamente da la clasificación de las empresas en sanas y quebradas, sino que también da la probabilidad de fracaso.

La misma metodología que para el análisis discriminante se usó para la aplicación del logit. Se construyeron dos modelos logit correspondientes a los atributos reducidos y a la tabla de información completa.

De acuerdo a los resultados, el modelo logit desarrollado a partir de la tabla de información completa da una mejor clasificación que el modelo

basado en los atributos de la tabla reducida. Comparado el modelo logit con el análisis discriminante, los dos modelos logit son superiores en todos los años, excepto en el año -1 para la muestra prueba. La comparación con la investigación rough set no es tan clara dado que este método es superior al otro para algunos años y viceversa.

Pero lo que queda claro es que la investigación rough set responde eficientemente a la predicción del fracaso empresarial, siendo además una buena alternativa tanto al análisis discriminante como al análisis logit.

El siguiente trabajo es el estudio de Greco *et al.*, (1997).

Utilizaron un método rough set presentado por Greco *et al.*, (1996) basado en aproximaciones de una partición dada de objetos en categorías previamente definidas y ordenadas, utilizando la relación de dominancia en lugar de la relación de no diferenciación para posteriormente aplicarlo a la predicción del fracaso empresarial. No vamos a comentar las diferencias con el enfoque tradicional rough set por que cae fuera de nuestra investigación y tampoco vamos a entrar en el comentario sobre los resultados obtenidos ya que no son comparables con el enfoque clásico. Simplemente queremos indicar que tanto la muestra como los atributos utilizados son los mismos que en el estudio de Slowinski y Zopounidis (1995).

Finalmente, el último trabajo que vamos a comentar es el de McKee, (2000).

Las variables utilizadas en este estudio son:

$A_1 = \text{Beneficio Neto} / \text{Activo Total}$

$A_2 = \text{Activo circulante} / \text{Activo Total}$

$A_3 = \text{Activo circulante} / \text{Pasivo circulante}$

$A_4 = \text{Cash} / \text{Activo Total}$

$A_5 = \text{Activo circulante} / \text{Ventas}$

$A_6 = \text{Deuda a largo plazo} / \text{Activo Total}$

$A_7 = \text{Existencias} / \text{coste de los bienes vendidos}$

$A_8 = \text{Cuentas a cobrar} / \text{Ventas}$

La selección de los ratios financieros se basó en su utilidad en estudios predictivos previos tanto del mismo autor como de otros. De estos ocho ratios un estudio previo del autor (McKee, 1995) emparejaba la partición recursiva con la teoría de la continuidad para reducir el conjunto de estas ocho variables predictivas a únicamente dos. Estos dos ratios eran el ratio del circulante (Activo circulante/ Pasivo circulante) y el ratio Beneficio Neto/ Activo Total. De esta forma en este estudio se compara la capacidad predictiva del enfoque Rough Set con el modelo de partición recursiva utilizado en el estudio previo.

En cuanto a los datos, se seleccionaron 200 compañías al azar (100 sanas y 100 quebradas) para los años 86-88 pertenecientes a muy diversos sectores, y con un periodo mínimo de existencia de cinco años.

De las 200 compañías, sólo 100 (50 sanas y 50 quebradas) se introdujeron en el RoughDAS (Slowinski y Stefanowski, 1994a) con los valores para los dos ratios. Los ratios financieros fueron discretizados utilizando 10 subintervalos que fueron determinados a través de los percentiles. La tabla de normas se introdujo en el RoughDas que transformó automáticamente los valores de las empresas para los dos ratios.

Dado que solo se utilizaron dos ratios no fue necesario efectuar la reducción de atributos, sino que directamente se obtuvieron 27 reglas de decisión (21 determinísticas y 6 no determinísticas) con una fuerza que variaba de 1 a 11 (media 3.96).

Las reglas obtenidas se utilizaron tanto para reclasificar la muestra de la cual se habían obtenida como las otras 100 compañías no utilizadas para estimar las reglas. Los cálculos se realizaron con el Rough Class (Mienko et al., 1995).

Para las empresas que encajaban con varias reglas que indicaban diferentes clases se utilizó la fuerza para asignarlas a la clase que indicaba la regla más fuerte. Para las compañías que no encajaban con ninguna regla se utilizó la VCR (véase capítulo 1) situando los niveles de indiferencia en 0, estricta diferencia en 1 y veto en 3. Los resultados obtenidos fueron muy buenos: para la muestra con la que se estimaron las reglas se obtuvieron 93 aciertos y 7 errores (2 clasificaciones erróneas y 5 empresas sin clasificar) y para la muestra holdout fueron 88 aciertos (2 clasificaciones erróneas y 10 empresas sin clasificar).

Los resultados fueron comparados con el modelo de partición recursiva que ya hemos mencionado introduciendo las doscientas compañías en dicho modelo. La tasa de aciertos bajó al 65%, con lo que se demostró que la superioridad del modelo Rough Set.

## **Capítulo 3**

# **SISTEMA ROSE : EXPLORADOR DE DATOS ROUGH SET**

## I.- ANTECEDENTES DEL ROSE Y OTROS ALGORITMOS EXISTENTES

La teoría rough set es una herramienta útil para el análisis de los datos de un sistema de información. En este análisis las partes esenciales son: la evaluación de la importancia de los atributos para la clasificación de objetos, reducción de los atributos y objetos superfluos, descubrir las relaciones más significativas entre los atributos de condición y la asignación de los objetos a las clases de decisión, y representación de estas relaciones en forma de reglas de decisión.

Una vez que los datos se han recogido, otra cuestión importante es la aplicación de los resultados obtenidos a la clasificación de *nuevos objetos* basándose en la experiencia previa.

Desde el punto de vista práctico, el análisis rough set de datos reales debería realizarse por un software en un ordenador, excepto cuando manejamos pequeños conjuntos de datos.

El programa para microcomputadoras ROSE (Rough Set Data Explorer ) es el software que hemos utilizado para la aplicación con datos reales de la teoría Rough Set. Sin embargo antes de entrar en su descripción comentaremos otros sistemas desarrollados para la aplicación en ordenador de esta teoría.

### *LEERS - Learning from Examples* (Grzymala-Busse, 1992)

Tal y como se ha mencionado el enfoque rough set, está especialmente indicado para tratar las inconsistencias de los datos; no corrige las inconsistencias que se presentan en los ejemplos o en los casos sino que calcula aproximaciones por arriba y por abajo. En consecuencia, y basándose en estas aproximaciones, se inducen dos conjuntos de reglas: determinísticas (ciertas) o no-determinísticas (posibles). Las reglas ciertas son categóricas y pueden utilizarse usando la lógica clásica. Las reglas posibles pueden posteriormente procesarse utilizando tanto la lógica clásica como cualquier otra teoría que trate la



incertidumbre. Un sistema pionero, y del que ya se ha hablado brevemente en la sección 1.7, en la inducción de reglas, tanto ciertas como posibles, se denominó LERS (en inglés, Learning from Examples based on Rough Sets). LERS trata las inconsistencias que existen en los datos de entrada mediante la utilización de los principios del enfoque rough set y la inducción tanto de reglas ciertas como posibles, lo hace sobre la base de aproximaciones por encima y por debajo para cada concepto. La primera implantación del sistema se hizo en 1988 y vamos a comentar brevemente algunas de sus características

El sistema LERS verifica la inconsistencia de los datos de entrada. En caso de que haya alguna inconsistencia en los mismos el sistema calcula las aproximaciones por encima y por debajo para cada concepto. El usuario tiene la posibilidad de elegir entre dos alternativas: la ingeniería del conocimiento y la adquisición de conocimiento. En el primer caso, el sistema induce una simple descripción discriminante mínima para cada concepto. Prácticamente eso significa, que el sistema induce un conjunto de reglas suficiente, que describen completamente cada concepto, aunque solamente algunos pares atributo-valor estén implicados en las reglas. Hay que advertir que en los datos de entrada puede existir mucho más conocimiento no descubierto por este enfoque. Si el usuario quiere descubrir más reglas debería utilizar la vía de adquisición del conocimiento. En este segundo caso, el sistema induce todas las reglas, cada una en forma mínima, que puedan derivarse de los datos de entrada. Este segundo camino es utilizado, por ejemplo, para sistemas expertos, cuando es crucial conocer todo cuanto se pueda acerca del problema que se esté considerando (inducción orientada al descubrimiento). Sin embargo la complejidad en términos de tiempo para la primera opción es polinómica, mientras que para la segunda opción es exponencial. De esta forma cuando tenemos una gran cantidad de datos de entrada, la segunda opción puede ser inviable.

En ambos casos, el usuario tiene la oportunidad de elegir entre un enfoque local o un enfoque global. En el enfoque local, un subsistema denominado LEM (en inglés, Learning from Examples Module) calcula la descripción mínima que sea discriminante, es decir, se seleccionan aquellos pares valor-atributo sobre la base de su relevancia para cada concepto. En el enfoque global, cada concepto es representado por una doble partición que le sustituye. Esta partición tiene dos bloques: el primer bloque es el concepto y el segundo su complementario. Se denomina *cobertura global* al subconjunto mínimo del conjunto de todos los atributos tal que de él dependa la partición sustitutiva.

El sistema de conocimiento LERS se desarrolló en la Universidad de Kansas.

#### *Método Rough Sets de Precisión Variable*

La idea de comentar esta variante del modelo original introducido por Pawlak es debido a que está incorporada en el ROSE. Además en la sección 2.1. se comenta una aplicación del mismo.

El modelo de Pawlak trata el análisis de dependencias de los datos determinísticas. Es decir, en su formalismo no reconoce la posibilidad de presencia o ausencia de relaciones no determinísticas, esto es, las que pueden llevar a reglas predictivas con probabilidades menores que uno. En algunos conjuntos de datos, sin embargo, la información disponible no es suficiente para producir reglas determinísticas fuertes pero sí es suficiente para determinar reglas fuertes pero no determinísticas que estimarían probabilidades de decisión. Para tratar este problema Ziarko (1993a) propuso el modelo rough set de *Precisión Variable*. Este modelo se implantó en el sistema DATALOGIC/R y los conceptos básicos que utiliza son dos:

- a) *Aproximación del espacio*: Esta aproximación tiene dos componentes, el universo objeto de consideración, que hemos

ido denotando por  $U$ , y una relación de equivalencia  $R$  que parte el universo en clases disjuntas denominadas conjuntos elementales.

La relación de equivalencia representa nuestra clasificación del conocimiento, es decir, nuestra capacidad de discernir objetos diferentes dentro del universo  $U$ . Dos objetos con los mismos atributos no son diferenciables lo cual conduce a la división de los objetos en clases de objetos idénticos (en el sentido de no ser distinguibles en base a unos atributos) o, lo que es lo mismo, en conjuntos elementales. El tamaño de un conjunto elemental nos indica el grado de precisión, o la granularidad de la representación del conocimiento.

Dicho grado puede ser controlado en DATALOGIC/R mediante la variación del parámetro "roughness" (precisión) en el intervalo,  $\langle 0,1 \rangle$ . La clasificación más "fina" se logra situándole en 0 y la menos precisa situándole en 1. Con un nivel bajo el sistema tiende a descubrir muchas reglas débiles (en el sentido que ya hemos comentado) pero con una probabilidad de decisión estimada alta. Estas reglas probablemente suelen ser incorrectas debido a la poca evidencia que las soporta. Cuando el parámetro se sitúa a niveles altos, el sistema identifica reglas más fuertes y más sencillas pero con probabilidades más bajas. Estas reglas son probablemente correctas o casi correctas pero su utilidad depende de la magnitud de la probabilidad del resultado que predicen.

- b) *Aproximación por abajo BETA de un conjunto* (un concepto)  $X$ , en el sentido de aproximación del espacio, es la unión de todos los conjuntos elementales cuyo grado de solapamiento (esto es, el tamaño de la intersección) con el conjunto  $X$  es menor o igual a BETA. Dado que las reglas obtenidas, se basan en identificar la aproximación por debajo, el parámetro BETA en la práctica

representa el mínimo admisible o la probabilidad de que la regla sea aceptable por el usuario. El usuario del sistema DATALOGIC/R puede controlar el nivel BETA en la búsqueda de las mejores reglas de compromiso, es decir, aquellas tanto con fuerza como con probabilidad relativamente altas.

#### *RoughDAS, Sensitivity Analyser y RoughCLASS*

Son los predecesores directos del ROSE y los vamos a describir con un poco de detalle ya que dicha descripción nos servirá para el ROSE.

RoughDas-Rough Sets based Data Analysis System- (Mienko *et al.* 1996a), es un programa para ordenadores personales creado para el análisis de sistemas de información basado en los rough sets. Permitía desarrollar los principales pasos de dicho método:

- la aproximación de cada clase de objetos (átomos),
- el cálculo de las aproximaciones y la calidad de la clasificación, y búsqueda de los reductos de los atributos,
- búsqueda, si existe, del núcleo (incluía la posibilidad de cuando existían muchos reductos, y era difícil elegir uno, ir añadiendo atributos al núcleo; en cada paso el usuario era informado sobre cada posible incremento de la calidad de la clasificación con el aumento de un nuevo atributo; se podía en definitiva, construir el reducto según las preferencias del usuario),
- reducción de los atributos no significativos del sistema de información,
- derivación del algoritmo de decisión del sistema de información reducido. Dicho algoritmo podía desarrollarse de manera automática o de modo interactivo (en el que el usuario controla el proceso de construcción de reglas seleccionando los atributos).

Se instaló como software interactivo en ordenadores IBM y bajo el sistema operativo DOS.

Ha sido utilizado con éxito en numerosas aplicaciones prácticas como hemos indicado en la sección 2.1.

RoughDas poseía sus propio ficheros, tanto los ficheros de datos como aquellos que se iban generando conforme se iban realizando las distintas etapas de la metodología rough set (sistema de información original y codificado, tabla de normas, reductos, núcleo, algoritmo de decisión).

Aceptaba datos en la forma de un sistema de información. Dado que los datos podían tener atributos cuantitativos y cualitativos, debían ser previamente codificados utilizando algunas *normas* (de las que ya hemos hablado en las aplicaciones, ver capítulo 2).

RoughDAS incorporaba un software para discretizar las variables continuas en discretas. La definición de los valores frontera podía, sin embargo, influir en la calidad de la clasificación. Por tanto era necesario analizar la sensibilidad de los resultados a los cambios en la definición de estos valores y, para ello, se creó un programa para analizar la sensibilidad frente a los cambios, *Sensitivity Analyser*. Este programa permite variar los extremos de los intervalos, tanto por la derecha como por la izquierda, y ver como influía dicha variación en la calidad de la clasificación. Solo permitía efectuar para un atributo cada vez. Además en cada punto de control el sistema calculaba la calidad de la clasificación.

Por otro lado para clasificar nuevos objetos, se busca, como se ha explicado anteriormente, desarrollar un algoritmo de decisión para encontrar una regla cuya parte de la condición sea satisfecha por la descripción de un nuevo objeto. Cuando se encuentra dicha regla, se utiliza para ayudar a la clasificación. En algunos casos, cuando no existe una regla adecuada en el algoritmo, se hace necesario encontrar las

reglas “más cercanas” (de acuerdo a una medida determinada de distancia). Dichas ideas se implantaron en el software denominado RoughClass.

RoughClass es un software creado para ayudar a la clasificación de nuevos objetos, es decir, aquellos objetos descritos únicamente por valores de atributos de condición.

El sistema analizaba el algoritmo de decisión para encontrar una regla cuya parte de la condición sea satisfecha por la descripción de un nuevo objeto. Si el sistema la encontraba se la presentaba al usuario, ya fuese determinística o no, y le informaba de la fuerza de las mismas. Si no existía tal regla, Roughclass, ayudaba al decisor presentándole las reglas más cercanas, según la “relación de proximidad valorada” (sección 1.9), indicándole la distancia a la regla, la fuerza de la misma y los objetos que la creaban.

Roughclass, al igual que RoughDAS, se instalaba en ordenadores IBM y sobre el sistema operativo DOS. Utilizaba la técnica del menú de windows y aceptaba los ficheros de datos del RoughDas. Es decir, su input de entrada consistía en el sistema de información codificado y en el algoritmo de decisión.

En RoughClass existe una opción para detectar y eliminar las posibles inconsistencias que pueden existir en el algoritmo de decisión.

Además incorporaba a la descripción de cada regla de decisión, el número de objetos en el sistema de información que encajaban con dicha regla, es decir, la *fuerza* de la misma.

## II.- ROSE – Software de aplicación de la Teoría Rough Set

Vamos a pasar a describir el software utilizado en nuestra aplicación, verificar todas las posibilidades que ofrece para poder aclarar en la parte empírica del trabajo cual ellas hemos seguido y por qué.

Para la descripción del ROSE vamos a seguir las indicaciones que los creadores del mismo ofrecen (Predki et al., 1998 y Predki y Wilk, 1999).

Queremos mencionar, antes de pasar a la descripción del programa, que cualquier ordenador personal conectado a internet puede acceder a la página web [www-idss.cs.put.poznan.pl/rose](http://www-idss.cs.put.poznan.pl/rose). donde se puede descargar el programa ROSE (se basa en la relación de no diferenciación), su manual e, informarse brevemente de algunas de sus características. Además la teoría Rough Set presenta variaciones desde su concepción clásica. Una de ellas está basada en la relación de dominancia en lugar de la no diferenciación. El software 4eMka incorpora la relación de dominancia y puede descargarse en la página web [www-idss-cs-put.poznan.pl/4emka](http://www-idss-cs-put.poznan.pl/4emka).

Dado que nuestro trabajo se basa en la teoría Rough Set en su versión clásica, nos limitaremos, a continuación, a la descripción del ROSE.

Hoy en día, el coste de adquirir y almacenar datos es relativamente bajo, por lo que frecuentemente se guarda información por medios electrónicos. Después de acumular la información los propietarios de los datos suelen querer analizarlos, persiguiendo todos o alguno de los siguientes objetivos:

- verificar la inconsistencia de los datos,
- reducir la información superflua,

- transformar los datos en conocimiento, esto es, descubrir patrones interesantes y útiles escondidos en los datos.

Existen actualmente, muchas técnicas que se pueden usar para conseguir los objetivos que se han descrito anteriormente (estadísticas, ingeniería del conocimiento,..). Uno de los métodos que pueden analizar datos de una tabla de información es la teoría rough set, que, tal y como hemos indicado, es especialmente útil cuando los datos presentan inconsistencias. Esta metodología es la que ha sido implantada en el software ROSE (Rogh Set Data Explorer).

El programa para microordenadores ROSE es un software modular de aplicación de esta teoría tanto en su enfoque clásico como con algunas nuevas extensiones basadas en el modelo de precisión variable o la relación de similitud, y de técnicas de descubrimiento de reglas. Fue creado en el *Laboratory of Intelligent Decision Support Systems* del Instituto de *Computing Science* en Poznan, basándose en catorce años de experiencia en el descubrimiento del conocimiento basado en el enfoque rough set y el análisis del proceso de toma de decisiones.

Funciona en PC's que utilicen sistemas operativos de la familia de 32 bit de Microsoft Windows (W95, W98 y NT). Es el sucesor del sistema RoughDas, que como hemos visto, es uno de las software más exitosos que incorpora la teoría rough set y que fue utilizado en muchas aplicaciones.

Sin embargo, y debido a las limitaciones del RoughDAS, especialmente su incapacidad para utilizarlo en los ordenadores actuales, se hizo necesario diseñar e instalar un nuevo software.

ROSE se ha diseñado para su fácil manejo utilizando un interface amigable y el sistema de dialogo de windows. Presenta, además, los resultados en este medio o a través de una hoja de cálculo. Todo esto le convierte en una herramienta sencilla para la exploración y análisis de los



datos. Esto trae como consecuencia que puede ser utilizado tanto por expertos como por usuarios ocasionales que quieran desarrollar una exploración sobre un conjunto de datos. El sistema se comunica con el usuario utilizando el dialogo de windows y todos los resultados se presentan en este medio. Además los datos pueden ser editados utilizando una hoja de cálculo como interface.

Está construido sobre una arquitectura modular, lo que significa que cada tarea se desarrolla por un módulo del programa autónomo. Luego, para facilitar su uso, todos los módulos se agregan en un único medio, el interface del usuario. Ya, en su origen, ROSE se creó como una serie de módulos independientes que posteriormente se unieron en un único sistema. Se mejoró su velocidad para permitir hacer análisis más rápidos sobre grandes conjuntos de datos y haciéndolo compatible con ordenadores con más potencia y sobre todo se creó un interface más amigable al utilizar como plataforma para el mismo el que ofrece Microsoft Windows.

Como consecuencia de todo lo anterior, el sistema ROSE tiene una importante característica y es que es relativamente fácil añadirle nuevos módulos lo que garantiza su mejora en el futuro.

Para facilitar su gestión el proceso de exploración del ROSE utiliza unas estructuras especiales denominadas *proyectos*. El proyecto contiene información relativa no solo a los ficheros de datos, si no también a los cálculos efectuados sobre los mismos.

ROSE acepta entrada de datos en forma de una tabla de información con todas las características que corresponden a este sistema de información y que ya hemos expuesto en otras secciones. Los datos se almacenan en ficheros de solo texto ya que garantiza una mejor reutilización de los mismos y su posible importación de bases de datos y de hojas de cálculo. Los ficheros de datos presentan un nuevo formato ISF (Information System File). Dicho formato permite una mejor revisión

de la sintaxis y una definición extendida de los atributos que pueden tener valores reales, codificados o incluso léxicos. Existe además un mecanismo de importación y exportación de datos a otros sistemas como RoughDAS, LERS, etc y también la posibilidad de obtener datos de los formatos de bases de datos más comunes (dBase o Paradox).

La especificación del fichero ISF tiene presente la posibilidad de nombres largos de atributos (por encima de 30 caracteres alfanuméricos) y valores condicionales de los atributos (tales como “alto”, “bajo”) además de valores reales y enteros. Dado que es un fichero de solo texto puede transferirse entre diferentes sistemas operativos sin modificaciones. Es además fácil de editar y de verificar la corrección de los datos contenidos en el fichero.

El formato de fichero tiene una estructura abierta. Se divide en secciones y se pueden añadir nuevas secciones hasta ahora indefinidas para un uso posterior. El usuario puede decidir ignorar alguno de los atributos tan solo cambiando la calificación de los mismos.

Excepto su visualización, todos los resultados se escriben en ficheros de solo texto, por tanto están preparados para sacarlos fuera del sistema y son fácilmente convertibles a otros formatos de ficheros.

El proceso de exploración de datos que utiliza el ROSE puede descomponerse en varias etapas.

En una primera etapa, el usuario del sistema ha de convertir los datos en una tabla de información y guardarlos en un fichero ISF. Para facilitar esta fase existe un mecanismo que importa datos de varias fuentes.

Después, en el caso de que falten algunos valores el usuario debe realizar una fase de reprocesamiento. Esta fase también es necesaria cuando existen atributos de tipo continuo o cuando posteriormente se va a utilizar la relación de no diferenciación en el proceso exploratorio.

Luego, el usuario tiene que seleccionar el modelo de exploración de datos. Actualmente el ROSE contiene tres modelos independientes que describiremos posteriormente. La selección del modelo puede implicar la indisponibilidad de algunas de las fases del proceso de análisis de datos.

El usuario también puede seleccionar los siguientes niveles de exploración de datos: aproximaciones rough, reducción de atributos, inducción de reglas y clasificación.

Vamos a describir con más detenimiento cada una de estas fases:

#### *Reprocesamiento*

Su finalidad es preparar los datos para un análisis posterior y convertirlos en el formato adecuado para utilizar el método que seleccionemos.

En las aplicaciones del mundo real suelen faltar valores al recoger un conjunto de datos. Ello supone un problema para la exploración de los mismos. El método incluye un módulo *completar* basado en el análisis estadístico de los valores más frecuentes (utilizado normalmente en la ingeniería del conocimiento).

Cuando el usuario selecciona el modelo de exploración basado en la relación de no diferenciación se sugiere reemplazar el dominio de los atributos continuos en un dominio de tipo discreto. A dicho proceso se le denomina discretizar y para lograrlo existen dos enfoques: una discretización basada en el conocimiento del experto y otra de manera automática basada en la teoría de la información. El sistema ROSE contiene ambos métodos. Es necesario especificar que la discretización de tipo automático utiliza un algoritmo basado en la medida de la entropía (Fayyad e Irani, 1992). Utilizando la discretización dirigida tenemos la posibilidad de crear una tabla de discretización o visualizar el dominio de los atributos en intervalos.

### *Métodos Disponibles*

Actualmente, hay tres modelos disponibles de exploración de datos en el sistema ROSE: el modelo clásico rough set, el modelo rough set de precisión variable y el modelo rough set basado en la similitud.

Desde la introducción en 1982 por Pawlak de la teoría rough set, ésta ha sido utilizada en muchas aplicaciones pero, sin embargo, hay conjuntos de datos para los que el modelo clásico se queda insuficiente, por lo que se hizo necesario recurrir a extensiones de esta teoría. Una de estas extensiones es el modelo de precisión creado por Ziarko (1993) en el que se cambia la descripción de las aproximaciones. Este método es especialmente útil en el análisis de conjuntos de datos con grandes regiones frontera. Otro enfoque diferente es el utilizado en modelo de relación de similitud introducido por Slowinski y Vanderpooten (1997). La similitud sustituye a la relación de no diferenciación y mantiene la propiedad reflexiva como única propiedad de entre las tres (reflexiva, simétrica y transitiva) que caracterizan a la no diferenciación. Permite el análisis de atributos numéricos sin una previa discretización y también induce reglas con una sintaxis que utiliza la relación de similitud.

### *Aproximaciones*

La parte más importante de la teoría rough set es la aproximación. El primer paso para encontrar las aproximaciones es crear conjuntos elementales (también denominados, como hemos visto, átomos). Un conjunto elemental contiene objetos que no son diferenciables en base a los atributos de condición, esto significa que estos objetos tienen idénticos valores para todos los atributos de condición. Cuando se utiliza el modelo de relación de similitud, los conjuntos elementales se sustituyen por clases de similitud.

Cuando el número cardinal de un conjunto elemental es más de uno es probable, que sus objetos pertenezcan a diferentes clases de decisión,

en este caso existe una ambigüedad. Este es el motivo por el que definimos dos aproximaciones:

a) La aproximación por abajo respecto al modelo de exploración que utilizemos es:

- en el modelo clásico – contiene todos los conjuntos elementales incluidos en la clase de decisión,
- en el modelo de precisión variable – contiene todos los conjuntos elementales que tienen al menos  $\beta * 100\%$  de objetos que pertenecen a la clase de decisión,
- en el modelo de relación de similitud – contiene todos los objetos cuyas clases inversas de similitud están incluidas en la clase de decisión,

b) La aproximación por arriba respecto al modelo de exploración que utilizemos es:

- en el modelo clásico – contiene todos los conjuntos elementales cuya intersección con la clase de decisión no da el conjunto vacío,
- en el modelo de precisión variable – contiene todos los conjuntos elementales que tienen al menos  $\alpha * 100\%$  de objetos que pertenecen a la clase de decisión,
- en el modelo de relación de similitud - contiene todos los objetos cuyas clases inversas de similitud tienen una intersección con la clase de decisión que no da el conjunto vacío.

ROSE incluye todas estas aproximaciones. Basándonos en estas aproximaciones se calcula la precisión de la aproximación, la precisión de la clasificación, y lo más importante de todo en la teoría rough set, la calidad de la clasificación.

### Reducción

Una de las propiedades más importante, según se ha mencionado, de la teoría rough set es la reducción de atributos. Tratamos de ver si existen algunos atributos redundantes en el sistema de información. Nuestro objetivo es buscar el mínimo conjunto de atributos que mantenga la misma calidad de clasificación que el conjunto completo de los atributos (esto significa que aproximan los datos de la misma manera). Ya hemos señalado que dichos subconjuntos se denominan reductos y el conjunto de los más significativos se denomina núcleo ( o parte común de todos los reductos).

Dado que el problema de encontrar reductos de una tabla de información dada puede ser complicado, es necesario desarrollar métodos que, si es posible encuentren dichos reductos en un tiempo razonable o introducir enfoques heurísticos que generen algunos reductos.

ROSE está actualmente equipado con cuatro métodos para generar reductos.

Históricamente el primer algoritmo introducido fue el de Romanski (1988) que intenta reducir la búsqueda en el espacio eliminando algunas partes que no tienen potencial como para ser incluidas en un reducto. Es útil cuando el número de reductos es relativamente pequeño (menos de 1000), debido a los requisitos de memoria.

En nuestros días el algoritmo más eficiente para la obtención de reductos es el desarrollado por Skowron *et al.* (1991). Es muy rápido aunque el coste de construir la matriz puede ser un factor desfavorable para aquellas bases de datos que tengan únicamente un par de reductos.

Para grandes bases de datos es posible buscar algunos reductos utilizando un método heurístico. Contiene una estrategia que se basa en ir

añadiendo atributos al núcleo. Es útil solamente cuando fallan otros métodos.

Finalmente, la última opción del ROSE en relación con los reductos es la generación manual de los mismos. El conjunto de atributos se presenta al usuario junto con el posible incremento o decremento de la calidad de la clasificación y que él sea quien decida que atributos añadir o quitar del conjunto. Este enfoque está especialmente indicado para expertos que tienen experiencia y conocimiento previo sobre el significado de las posibles coaliciones entre los atributos.

También por supuesto está la opción de generar el núcleo.

#### *Inducción de Reglas*

El conocimiento analizado en un conjunto de datos puede expresarse en forma de reglas de sentencias lógicas del tipo “si...entonces”. Anteriormente se ha explicado como una regla de decisión consiste en una parte de condición (conjunción de algunos resultados elementales sobre los valores de los atributos) y una parte de decisión (asignación a una o más clases de decisión).

La inducción de reglas es un problema autónomo que puede considerarse de manera independiente de la teoría Rough Set. Los algoritmos de inducción generan reglas para un conjunto dado de objetos. En el caso más sencillo tal conjunto consiste en todos los objetos de una determinada clase de decisión. La metodología Rough Set es especialmente útil cuando el conjunto de datos es inconsistente y objetos descritos por los mismos valores de condición son asignados a diferentes clases de decisión. En esta situación las reglas de decisión pueden generarse de las aproximaciones o de las fronteras de las clases de decisión.

El usuario puede elegir alguno de los tres esquemas de inducción de reglas:

- 1- Descripción mínima: la descripción resultante es un conjunto mínimo de reglas (es decir, el menor conjunto de reglas) que cubre a todos los objetos de un conjunto dado (una regla como hemos visto, cubre un objeto, cuando todas las condiciones en la parte de condición de la regla son ciertas para los valores de los atributos del objeto).

Esta descripción se genera utilizando el algoritmo LEM2. (Grzymala-Busse, 1992). Dependiendo de la definición del conjunto de objetos, para los cuales estamos generando las reglas, LEM2 induce dos tipos de reglas: *reglas exactas* (que son las generadas por el conjunto de objetos definidos como la aproximación por debajo de una clase de decisión dada) y *reglas aproximadas* (que son las generadas por el conjunto de objetos definidos como frontera de una clase de decisión).

Además del LEM2 original, ROSE contiene dos versiones modificadas del mismo: el algoritmo LEM2 con extensión de intervalos y el LEM2 con la extensión de similitud.

- 2- Descripción satisfactoria: la descripción resultante contiene únicamente reglas que satisfacen los requisitos especificados por el usuario (por ejemplo reglas que tengan una determinada fuerza o que tengan buenas posibilidades de discriminación).

Dentro de este esquema ROSE contiene el algoritmo desarrollado por Mienko *et al.* (1996a). Este algoritmo comienza con la generación de las reglas más cortas (que contienen una condición en su parte condicional), y gradualmente incrementa la longitud de la regla generada.

El espacio de búsqueda es limitado por los niveles definidos por el usuario: *longitud máxima de la regla* (es decir, el máximo número de condiciones en la parte de condición de la regla),



*fuerza mínima de la regla* (es decir, el número mínimo de objetos cubiertos por la regla, que pertenecen a la clase de decisión señalada por la regla) y *nivel mínimo de discriminación* (esto es, el ratio del número de objetos cubiertos por la regla, que pertenecen a la clase de decisión indicada por la regla, con respecto al número de todos los objetos cubiertos por la regla).

- 3- Descripción exhaustiva: La descripción generada contiene todas las reglas posibles que pueden inducirse del conjunto de objetos dado. Para lograr esto, la longitud máxima debería ser igual al número de atributos y la mínima fuerza debería ser situada en uno. Hay que resaltar, sin embargo, que la generación exhaustiva de todas las reglas puede ser extremadamente costosa en términos de tiempo y de utilización de memoria, incluso para conjuntos de datos de tamaño medio.

#### *Clasificación*

En esta fase, las reglas de decisión generadas en el paso previo se utilizan para clasificar objetos (o sea, para asignarlos a una clase de decisión). Es decir, ROSE también puede usarse como un tipo de sistema de apoyo a la clasificación basado en un conjunto de reglas. El proceso de asignación se desarrolla en las siguientes etapas:

- 1- si un objeto es cubierto exactamente por una regla, entonces el objeto es asignado a la clase de decisión indicada por la regla.
- 2- si hay varias reglas que cubren un objeto señalando a diferentes clases de decisión, el conflicto se resuelve asignando el objeto a la clase con el mayor número de votos. Para ello el programa nos pide el parámetro denominado *umbral de mayoría*. Por ejemplo, si un objeto encaja con las siguientes reglas:

R1- clase =1 fuerza: 2

R2- clase =2 fuerza: 4

R3- clase =3 fuerza: 1

R4- clase =1 fuerza: 3

Los votos se calculan para cada clase de decisión como suma de todas las fuerzas de las reglas de decisión que señalan dicha clase. En nuestro ejemplo:

Clase 1: 5 votos (regla 1+ regla 4)

Clase 2: 4 votos

Clase 3: 1 voto

Número total de votos: 10

Los votos en valores relativos serían:

Clase 1: 50%

Clase 2: 40%

Clase 3: 10%

Escogeríamos la clase con mayor número de votos en términos relativos (si la elección no es obvia se haría de manera aleatoria) y compararíamos dicho valor con el umbral de mayoría. Si el número de votos es mayor o igual a dicho umbral, el objeto se clasifica en la clase seleccionada. En cualquier otro caso el *clasificador* no puede tomar una decisión y la decisión final generada por el mismo es *NINGUNA*.

3- si ninguna regla cubre un objeto, el objeto se asigna a la clase de decisión indicada por la regla más cercana. Es decir, asignamos el objeto utilizando reglas que encajan parcialmente (reglas que al menos una condición encaja con la descripción del objeto)

En este caso el programa nos pide dos parámetros: *mínima semejanza* y *reglas que encajan parcialmente*. Dado que la semejanza entre las reglas y el objeto a clasificar es menor del

100%, estos parámetros se utilizan para acortar la distancia con el conjunto de estas reglas. La semejanza mínima indica cómo de similar debería ser la regla para poder usarla para clasificar el objeto (si situamos este parámetro en el 0%, se utilizarían todas las reglas incluso aquellas que no encajasen con ningún atributo de condición). Una vez que hemos obtenido las reglas más similares según el parámetro anterior, podemos utilizar el conjunto de todas ellas para clasificar el objeto o podemos realizar la clasificación sólo con las  $k$  reglas más similares (siendo  $k$  el número de reglas que encajan parcialmente).

Dicha similitud entre la regla y el objeto se calcula como una medida de distancia. Esta distancia es una métrica euclídea que se calcula únicamente para los atributos que coinciden tanto en el objeto como en la regla. La medida de distancia es la métrica  $L_p$  descrita en Slowinski (1993) y comentada anteriormente.

Si hay varias reglas con la mínima distancia al objeto, se utiliza la misma estrategia descrita en apartado anterior para resolver el conflicto.

La clasificación implantada en el ROSE se puede utilizar para dos tareas diferentes:

- a) Clasificación de un nuevo objeto, cuya pertenencia a una clase de decisión no sea conocida.
- b) Clasificación de un objeto ya clasificado (también denominada reclasificación).

La primera de las tareas mencionadas es llevada de forma interactiva para cada único objeto. Después de la clasificación, al usuario se le presenta la clase de decisión asignada y la información detallada que se ha considerado durante el proceso de clasificación.

La segunda tarea es un único paso formado por un test de reclasificación. Al usuario se le presentan los resultados después de desarrollar la totalidad del test. ROSE ofrece dos escenarios para dichos tests:

- “Dejar uno fuera” (“living one out”) – sugerido en el caso de pequeños conjuntos de datos (menos de 100 objetos). En cada pasada la precisión de la clasificación (definida como número de objetos correctamente clasificados sobre el número total de objetos a clasificar) es o 0% o 100% (es decir, se clasifica únicamente un objeto). Después de completar el test, únicamente se calcula la precisión media. En este caso la desviación típica no tiene ningún valor.
- “K-veces validación cruzada” (K-fold cross-validation) – se debe utilizar para conjuntos de datos mayores. El usuario puede elegir una división aleatoria o estratificada en varias veces (en la última división, la distribución del número de objetos de la clase de decisión en cada vez es la misma que para el conjunto de datos en su totalidad). En este caso después de  $k$ -veces obtendremos  $k$  precisiones, por tanto hay que calcular la media de todas ellas, que se presenta junto con la desviación típica para mostrar cómo de estable es la precisión. Por tanto, en este segundo caso, después del test se presenta una detallada información estadística acerca de la precisión de la clasificación (esto es, medias, desviaciones típicas, distribución en clases de decisión). Todo esto se complementa con una matriz de confusión.

## **Capítulo 4**

### **LAS VARIABLES DEL MODELO**

## **I.- LA VARIABLE DEPENDIENTE: CONCEPTO DE CRISIS EMPRESARIAL. INSOLVENCIA**

Sin ánimo de ser exhaustivos, queremos describir brevemente qué entendemos por crisis empresarial, su relación con la insolvencia, sus efectos y sus causas con la intención de poder justificar las variables (ratios) que vamos a introducir en el modelo que aplicaremos para la predicción de la misma.

La expresión crisis empresarial, según los autores que consideremos, está sujeta a distintas acepciones. En un primer momento nos referiremos a la crisis de carácter financiero fundamentalmente.

Esta concepción financiera de crisis empresarial es muy amplia. Esta amplitud tanto en su significado como en su utilización lleva a la ambigüedad, ya que el término fracaso o crisis empresarial se emplea tanto para referirse a la insolvencia técnica, cuando la empresa es incapaz de atender sus obligaciones financieras, o a la insolvencia definitiva, cuando el valor total de los activos de la empresa es menor que el valor de sus pasivos.

Por tanto, el concepto de solvencia se refiere normalmente a dos componentes relacionados pero distintos. El primer componente se refiere a la disponibilidad de medios líquidos necesarios para hacer frente a las obligaciones derivadas de la gestión corriente y de inversiones en una empresa puntualmente, y el segundo, a la capacidad para obtener beneficios y generar tesorería.

En consecuencia, nos estamos refiriendo no solo a la solvencia necesaria para un ejercicio, sino también a la garantía de solvencia (reflejada en una adecuada estructura patrimonial y una conveniente dimensión cuantitativa de fondos generados) que tiene una empresa de manera permanente.

Quizás, la voz insolvencia sea menos ambigua en su utilización, y se podría definir, en términos exclusivamente financieros, como la incapacidad de una entidad para pagar sus deudas.

Un primer periodo de insolvencia puede ser superado mediante aplazamientos o medidas de reestructuración empresarial, o por el contrario desembocar en una crisis definitiva que termine en venta o absorción de la empresa, en situaciones concursales, o en cierre definitivo.

La crisis empresarial es el colofón del proceso de agravación de la insolvencia, que culmina en el fracaso empresarial, excluyendo la venta o absorción, para entrar en procesos judiciales de quiebra o suspensión de pagos o en cierre empresarial.

Sin embargo, cabría preguntarse sobre cómo llega a una empresa a ser insolvente. La variedad de situaciones por las que puede transitar una empresa insolvente agrava aún más la ya de por sí difícil tarea de dar de manera unívoca y concreta una definición de *fracaso empresarial*. La realidad nos muestra cómo situaciones y procesos similares pueden terminar en soluciones de venta de una empresa, mientras que otra termina en cierre. También existen casos en los que la declaración por parte de los tribunales de una suspensión de pagos supone la confirmación del fracaso e inicio de la liquidación de la empresa, mientras que en otros casos, la empresa declarada en suspensión de pagos continúa sus actividades y consigue un acuerdo de quita o/y aplazamiento en el pago de las deudas, para seguir operando con normalidad.

Esta diversidad de situaciones nos obliga a definir un concepto de insolvencia para nuestro trabajo en función tanto de nuestros objetivos como de la disponibilidad de datos.

Nuestra definición de concepto de fracaso está basada en situaciones concursales ya que es un concepto riguroso, ajeno a interpretaciones diversas y presente en bases de datos asequibles, que supone un plus de objetividad para cualquier investigación empírica. En concreto, y dado que nuestra investigación se centra en el sector asegurador, una *empresa fracasada* será aquella que esté en una situación de dificultad grave, de desequilibrio patrimonial y financiero que afecta negativamente a la continuidad de la empresa. Por tanto, todas ellas han sido intervenidas por la C.L.E.A (Comisión Liquidadora de Entidades Aseguradoras). Se trata de una medida objetivamente determinable de las empresas que fracasan.

Brevemente señalaremos qué es la CLEA, cuáles son sus funciones y cómo es el proceso de liquidación en el que interviene la misma. (*Fuente página web Dirección General de Seguros*).

La Comisión Liquidadora de Entidades Aseguradoras es un organismo autónomo de la Administración General del Estado, vinculado al Ministerio de Economía cuyo control de eficacia corresponde a la Dirección General de Seguros y Fondos de Pensiones.

Sus funciones son las siguientes: Asumir la liquidación de entidades aseguradoras cuando le sean encomendadas por el Ministro de Economía o, en su caso, por el Órgano competente de la respectiva autoridad autonómica; actuar como Interventor único en los procedimientos de Suspensión de Pagos y como Comisario, Síndico y Depositario en los de quiebra.

La C.L.E.A. desarrolla el proceso de liquidación conforme a lo que establece la Ley 30/1995 según el siguiente esquema:



**A) Fase I. Máximo de 9 meses**

- Se inicia con la toma de posesión de la entidad. Incluye el cierre de las sucursales y la fijación de la sede de la liquidación.

- Se negocia el expediente de despido colectivo, fijando el número de personas necesarias para el desarrollo de la liquidación, procediéndose a la regulación de empleo del resto de la plantilla.

- Se publican anuncios en los periódicos y se remiten cartas a los asegurados informando de la liquidación, especialmente si se ha decretado el vencimiento anticipado de los contratos.

- Se abre el plazo para la presentación de las reclamaciones de los acreedores. Recepción de dichas reclamaciones. Se inventarían los bienes y se recoge la documentación social.

- Se formula el balance provisional que incluye los beneficios previstos en la Ley, conforme al cual el Consejo de Administración de la C.L.E.A. aprueba el porcentaje a abonar a los acreedores por contrato de seguro.

**B) Fase II. Duración entre 2 a 3 años dependiendo del volumen y de los ramos en que operaba la compañía en liquidación.**

Durante ese tiempo la Comisión adquiere los créditos a los acreedores por contrato de seguro conforme se vayan completando los expedientes y en el porcentaje que el Consejo de Administración haya aprobado.

Asimismo se enajenan los activos de la compañía y se interponen, en su caso, las acciones judiciales necesarias para exigir las responsabilidades a que hubiera habido lugar y para reintegrar al activo los bienes de la misma.

**C) Fase III. Su duración depende de si la ejecución del plan de liquidación (pagos a los acreedores) en un solo período (duración mínima 6 meses) o si la ejecución del plan de liquidación (pagos a los acreedores) es en dos períodos (indeterminado).**

La ejecución del plan de liquidación se produce en dos períodos, cuando encontrándose pendientes acciones judiciales para no perjudicar a los acreedores, se procede al pago del efectivo existente en el primer período, esperándose a repartir en el segundo, lo que se obtenga de dichas acciones judiciales pendientes.

Se inicia con la celebración de la Junta de Acreedores que aprueba el plan de liquidación. Una vez ratificado por la Dirección General de Seguros y Fondos de Pensiones, se abona al resto de los acreedores sus créditos en el porcentaje que haya correspondido según el activo y pasivo de la entidad. Posteriormente se publica el balance final y se extingue la compañía en el Registro Especial de Entidades Aseguradoras de la Dirección General de Seguros y Fondos de Pensiones y en el Registro Mercantil.

Por tanto, una vez visto cómo interviene la CLEA, en este estudio al perseguir un análisis de la situación financiera de las empresas, donde pretendemos discriminar entre empresas sanas frente a empresas que no lo están, utilizaremos el concepto de fracaso en un sentido jurídico (empresa intervenida por la CLEA) como indicador de situación de crisis en la empresa.

La definición de este concepto es un factor que influye con fuerza en los resultados obtenidos, porque las empresas que aún estando en situación de insolvencia extrema son vendidas, no están en el bloque de las fracasadas, al mismo tiempo que otras en situación legal de quiebra o suspensión de pagos, y por tanto se encuentran en el bloque de las quebradas, continúan sus actividades y consecuentemente son empresas vivas.

Por esta razón es importante considerar que las técnicas de clasificación y predicción no pueden ser eficientes al cien por cien por la propia naturaleza del problema, por lo que en consecuencia con lo dicho cuando se cuantifica un porcentaje de aciertos debe añadirse un coeficiente de sesgo a dicha cifra (Sanchis, 2000).

## **II.- LAS VARIABLES DEPENDIENTES: LOS ESTADOS FINANCIEROS COMO INSTRUMENTOS DE LA PREDICCIÓN DE LAS CRISIS EMPRESARIALES. LOS RATIOS**

### **II.1.- Introducción: Enfoque Metodológico seguido en la selección de los ratios.**

En línea con el planteamiento financiero de crisis empresarial, y de acuerdo con nuestro objetivo de predicción, debemos centrarnos en dos hechos:

- a- En primer lugar, debemos efectuar un análisis de las causas, como paso previo y fundamental para intentar detectar las señales tempranas de alerta de posibles insolvencias para que nos ayuden a anticiparnos.
- b- Y, en segundo lugar, dado que estos indicios suelen tener su reflejo en la información financiera, hacer un análisis de la misma al objeto de seleccionar aquellas variables que puedan ser introducidas en el modelo.

Las causas del fracaso de las empresas son muy diversas y generalmente son varias las que concurren para llegar a una situación de crisis. En ocasiones el origen del problema es externo (indicios del entorno); también puede ser interno (indicios internos) y a menudo influyen ambos tipos, aunque siempre hay que señalar que con independencia del número y origen de los problemas el equipo directivo de la empresa tiene la responsabilidad de abordarlos y encontrar soluciones.

Algunas de las causas de fracaso empresarial son las siguientes:

*Causas externas a la empresa o derivadas del entorno:* Podríamos citar muchas, pero a modo de ilustración tendríamos: cambios en el mercado

(cambios en la estructura competitiva, en el nivel de precios); cambios en la tecnología, producción marketing y distribución; cambios económicos, culturales y sociales radicales y significativos (cambio de gobierno, guerras, fase depresiva del ciclo...).

*Causas internas a la empresa y que indican el deterioro de su situación:* Podríamos citar a modo de ejemplo las siguientes: mala gestión de la dirección, estrategias erróneas e inadecuadas, pérdida de productividad, inversiones improductivas, excesivo endeudamiento (agravado por altos tipos de interés) y deterioro de la estructura del pasivo, final del ciclo de la vida de un producto, fracaso de empresas del grupo, crecimiento de los impagados, etc.

Además, hay que tener en cuenta algunas situaciones especiales, como lo que ocurre con las empresas de nueva creación, las cuales presentan una tasa de mortalidad muy elevada en los primeros años de actividad.

De todo lo expuesto, lo que resulta innegable es que las causas o factores del fracaso empresarial son muy numerosas y/o variadas. Pero, estas señales pueden ayudarnos a predecir, en primer lugar, para, en segundo lugar, prevenir. Luego es necesario analizarlas cuando se presentan.

Como resultado, y desde nuestra mentalidad matemática, tenemos que es muy difícil establecer una relación funcional tipo causa-efecto, que en principio puede parecer el enfoque más directo.

Esto nos conduce (al igual que en otros muchos trabajos de investigación) a abandonar las causas para centrarnos en descubrir las consecuencias de los problemas indicados en la información económica y financiera de las empresas. Nuestra intención es detectar estos indicios o

indicadores en dicha información para posteriormente inferir la capacidad de uno o más de estos indicadores para explicar o predecir el fracaso empresarial.

Por tanto, el fracaso empresarial aparece como resultado de un proceso económico complejo con múltiples causas y cuyas consecuencias tienen su reflejo en los estados financieros presentados por las sociedades, (balance, cuenta de resultados y memoria), por lo que, tanto el análisis interno como el externo, pueden contribuir a diagnosticar anticipadamente el riesgo de fracaso de una empresa. Constituyen, por tanto, una herramienta esencial para los distintos usuarios de la información financiera.

Pero, hay que añadir otra dificultad a las mencionadas anteriormente, y es que la predicción está envuelta en una gran incertidumbre derivada, en este caso y según lo expuesto en el párrafo precedente, de factores tales como la inexistencia de información fiable y oportuna sobre la situación y perspectivas de la empresa y el sector, o que existan acontecimientos imprevisibles que lleven a que la quiebra sobrevenga de manera indetectable y sin reflejo en la información financiera (Jiménez *et al.*, 2000).

Todo esta complejidad que estamos relatando, lleva a que no exista como tal una “Teoría del fracaso empresarial”, por lo que las investigaciones acuden al terreno de lo empírico para ver qué efectos tienen en la información financiera las situaciones de insolvencia. Y, todo ello, y de acuerdo con lo que hemos indicado, considerando que el deterioro financiero no sobrevenga de manera repentina sino que se produzca de manera gradual que haga que la empresa entre en una situación cada vez más grave.

Si se efectúa un primer análisis de la información financiera relativa a las empresas fracasadas, nos encontramos una serie de variables que están relacionadas con la insolvencia: pérdidas, flujos de caja negativos,

endeudamiento alto o mal estructurado y que conduce a una carga financiera excesiva, fondos propios insuficientes, fondo de maniobra negativo, incumplimiento o no-renovación de deudas al vencimiento, etc.

Sin embargo, una vez más, y tal y como hemos señalado, en muchos casos observamos estas situaciones en las empresas fracasadas lo que nos resolvería el problema sin más que observar estas variables; pero, el problema es más complejo, ya que estos síntomas o indicios también están presentes en empresas que luego no fracasan.

Lo que nadie pone en duda es que estas variables nos informan con seguridad de problemas presentes y futuros, pero no basta con ello.

En consecuencia, habría que incorporar al análisis el grado de intensidad del problema. En este nivel de intensidad estaría una primera explicación de por qué hay muchas empresas que tienen alguno de los problemas citados, con carácter circunstancial o en períodos largos, e incluso de forma permanente (fondo de maniobra negativo de los grandes almacenes), sin que peligre su futuro.

Por tanto, para caracterizar las empresas desde el punto de vista de la predicción de la insolvencia, es necesario identificar las variables *más* afectadas por la situación y su cuantificación de forma que resulten discriminantes para establecer qué comportamiento es propio de las empresas solventes y no de las insolventes y al contrario.

A todo lo expuesto hay que añadir otra problemática. Ya que nos centramos en las consecuencias en la información financiera de las crisis empresariales, ocurre que distintas causas pueden producir efectos también distintos en las variables financieras. Por ejemplo, un alto endeudamiento puede producir efectos en incumplimiento de compromiso de pago y también un mejor resultado derivado de un apalancamiento financiero positivo.

Hemos mencionado que, pese a la problemática descrita, el objetivo es *identificar* las variables y *cuantificarlas*.

Los estados financieros contables de las empresas se consideran como una de las más valiosas fuentes de datos para la previsión de las crisis. Su análisis nos permite lograr la primera vertiente de nuestro objetivo: identificar las variables.

A través de los estados contables se examina información económica y financiera que caracteriza a una empresa, para poner de manifiesto sus aspectos estructurales y sus movimientos financieros y, de esta manera, poder emitir un diagnóstico sobre ella. Sin embargo, aún cuando el análisis de los estados financieros permite obtener conclusiones sólidas sobre la rentabilidad y la solvencia de una empresa y algo más limitadas para la liquidez, no suelen ser suficientes para poder analizar adecuadamente la evolución.

En este sentido, no hay que olvidar que los estados financieros en general y la contabilidad en particular no nacieron ni están pensados con el objetivo básico de que un tercero pueda analizar adecuadamente la situación actual y mucho menos la futura. Son principalmente, el medio que tiene los administradores para rendir cuentas de su actuación ante terceros.

Pero una vez identificadas las variables en los estados financieros la forma que presenta más ventajas de tratar esas variables y cuantificarlas es mediante el cálculo de determinados *ratios financieros* en los que intervengan las mismas.

Antes de dar unas breves notas sobre los ratios financieros, queremos ya manifestar que cualquier modelo que solo utilice la información contenida en los estados financieros (máxime cuando se utilizan solo el balance y la cuenta de pérdidas y ganancias) presenta una serie de limitaciones. Pero, sin



embargo, estos datos que podríamos denominar *externos*, son datos disponibles para cualquier interesado, por ser datos de obligada publicidad. Con lo que es fácil desarrollar un modelo (aunque limitado) en función de los datos disponibles para poder mejorarlo mediante la incorporación de otra información, si se dispone, de carácter interno y/o no publicable y añadir dichas variables al modelo.

Por otro lado, y hablando de los ratios, la utilización de los mismos permite comparar con otros valores que, el mismo ratio, haya tomado en otros tiempos (análisis histórico o de tendencia) o con ciertos estándares propuestos, tales, como por ejemplo, la media del sector industrial (análisis normativo).

El análisis histórico o de tendencia lleva al estudio de las series temporales de ratios de una misma empresa cuyo objetivo suele ser la predicción del valor futuro de los mismos (análisis predictivo).

El análisis normativo al estudio comparativo entre empresas mediante la utilización de los ratios como estándares. Se trata del establecimiento de ciertos valores patrones con los cuales comparar el ratio de la empresa analizada.

Una ventaja que presenta la utilización de los ratios es que reducen la dispersión en las cifras de los estados financieros de las empresas derivada del tamaño de las mismas. Esto facilita la comparación interempresas. Las empresas de gran tamaño producen grandes números contables y las de pequeño tamaño números contables menores. Por tanto, la utilización de ratios presenta la ventaja de reducir el sesgo que podría producirse derivado del tamaño de las empresas.

Cuando realizamos un análisis de la crisis empresarial mediante ratios contables, esperamos que los ratios, de un modo u otro, se deterioren

conforme se aproxima la fecha de la quiebra y que dicho deterioro sea perceptible o medible en un intervalo de tiempo relativamente amplio, a veces de hasta cinco años (Calvo-Flores y García (coord.), 1998b)

Un deterioro de los ratios se produce:

- cuando los valores de los ratios de las empresas en dificultades se apartan, en la dirección incorrecta, de los valores estándares o normativos del sector que estemos considerando.
- cuando los valores de las empresas candidatas al fracaso se tornan inestables en el tiempo, sufriendo oscilaciones o altibajos medibles, por ejemplo, con la desviación típica de los valores que tomen.

Dicha inestabilidad puede verse inducida por cambios bruscos en la política de la empresa que se reflejan en los estados financieros y contables y también por las denominadas prácticas de contabilidad creativa.

En definitiva, los modelos para predecir el fracaso empresarial o la futura solvencia de la empresa, fundamentalmente utilizan como variables explicativas los ratios económicos y financieros (extraídos de la información contable).

Los modelos de predicción del fracaso empresarial, en relación con los ratios contables, cumplen una doble misión:

- a) prueban el contenido informativo de los datos contables sobre la solvencia de la empresa, al medir la relación entre los datos y el fracaso,
- b) y, en segundo lugar, derivado del anterior, son un mecanismo para pronosticar el fracaso empresarial. Constituyen una

herramienta útil para la toma de decisiones por parte de los múltiples usuarios del análisis de la solvencia.

La identificación de las variables o indicadores de la insolvencia permite estudiar su capacidad predictiva en el tema que nos ocupa, la predicción de quiebras o del fracaso empresarial.

Por tanto, identificadas las variables (los ratios), pretendemos conocer mediante la aplicación de un modelo (en nuestro caso el modelo rough set) la relación entre las variables financieras y la solvencia o insolvencia de las empresas estudiadas.

Queremos anticipar, que indicadores o variables financieras relacionados con la medición de la insolvencia son variados, habiéndose contrastado muchísimos. Es frecuente en numerosos estudios empíricos partir de una gran cantidad de ratios para que el modelo propuesto reduzca el número de los mismos.

En nuestro caso, el enfoque metodológico que hemos utilizado para encontrar la relación entre las causas o síntomas y situaciones de insolvencia crediticia o total ha consistido en seleccionar diversas variables consideradas a priori como apropiadas (para las cuales justificaremos su utilización), para contrastarlas utilizando el método rough set. Es decir, hemos considerado no introducir una gran cantidad de ratios sino previamente seleccionar una serie de ellos. Pensamos que introducir mucha información puede incidir negativamente en los objetivos propuestos, por lo que preferimos no introducir información superflua en el modelo, que, lejos de contribuir a los fines predictivos, los dificulte.

Este enfoque permite seleccionar las variables más significativas para explicar la situación en función de los datos manejados, así como identificar si las conclusiones obtenidas son eficientes también cuando se aplican a

otras empresas de la primera muestra estudiada, y a los datos correspondientes a las empresas uno, dos, tres o más años anteriores al del estudio inicial (año de fracaso).

## II.2.- Selección de la muestra.

La muestra utilizada es la seleccionada para la aplicación del análisis discriminante en la predicción de la insolvencia en entidades aseguradoras españolas no vida (Sanchis, 2000 y 2002). Consta, en el ámbito nacional, de 36 empresas sanas y 36 empresas fracasadas, emparejadas fundamentalmente por tamaño y sector, para minimizar el efecto de estas variables en el estudio sobre los ratios que haga que diferencias en alguno de ellos se deban a otros factores ajenos a su situación financiera.

Aunque para un mayor detalle sobre la selección de la muestra se puede consultar la bibliografía citada anteriormente, queremos resaltar brevemente las características de la muestra.

La definición de *empresa fracasada*, tal y como ya se ha indicado, es aquella empresa que ha tenido que ser intervenida por la C.L.E.A. Todas ellas son Sociedades Anónimas, por tanto se han excluido del análisis otro tipo de formas societarias (Mutuas, Cooperativas).

El tipo de muestreo ha sido por *emparejamiento* controlado por unos determinados factores. Con el *muestreo por parejas*, se pretende que las dos submuestras (empresas sanas y empresas fracasadas) tengan características similares para aquellos factores que no son objeto del estudio con la finalidad de que, cuando se obtengan los resultados de clasificación, se puedan atribuir estos al factor que queremos explicar: el fracaso de las empresas.

Por tanto, la muestra está formada por el mismo número de empresas fracasadas que sanas y son muestras no demasiado heterogéneas ya que el emparejamiento ha tenido en cuenta el factor tamaño (medido a través del volumen de primas) y el tipo de negocio.

La información utilizada en el planteamiento de este trabajo, ha sido eminentemente contable. Estos datos están recogidos en los estados financieros de las empresas, de los cuales sólo se ha podido obtener el Balance y la Cuenta de Pérdidas y Ganancias. Los datos están extraídos de la publicación anual de “Balances y cuentas. Seguros privados” de la Dirección General de Seguros.

La utilización de documentos contables permite obtener una información cuantificable, objetiva y sistemática de la realidad económica y financiera de la empresa.

Para ver el poder explicativo de los ratios, se han tomado los datos a lo largo de cinco años antes de la quiebra, tomando como año base el primer año anterior a la misma en el caso de las fracasadas, y por extensión, también para su pareja ese será el año base. Sin embargo, dado que en 1982 se introdujo una nueva estructura en las cuentas anuales, no se ha utilizado información anterior a esa fecha para que la misma fuese lo más homogénea posible. Por tanto la serie temporal abarca datos desde el 83 al 94.

Las empresas utilizadas en el análisis aparecen en las Tablas 4.1 y 4.2.

**TABLA 4.1: EMPRESAS INTERVENIDAS POR LA CLEA**

Nº	NOMBRE	CODIGO	AÑOS
1	Kairos, cía. De seguros y reaseg.	C-043	93-92-91-90-89
2	Igualatorio Médico Palentino de seguros	C-130	93-92-91-90-89
3	Asistencia Sanitaria 2000 <b>(1)</b>	C-454	93-92-91-90-89
4	Sociedad Andaluza de Seguros	C-507	93-92-91-90-89
5	Conseguir, S.A. de Seguros Generales	C-598	93-92-91-90-89
6	Unión social de Seguros	C-638	92-91-90-89-88
7	Mundi Seguros <b>(2)</b>	C-663	92-91-89
8	Apolo	C-008	91-90-89-88-87
9	Unión Europea de Seguros	C-568	91-90-89-88-87
10	Segurauto	C-573	91-90-89-88-87
11	Reunión Grupo 86	C-440	90-89-88-87-86
12	Servicios Médicos	C-450	90-89-88-87-86
13	Larra	C-561	90-89-88-87-86
14	Unasyr, Unión alicantina de Seguros	C-567	90-89-88-87-86
15	Mades Fondo Asegurador <b>(3)</b>	C-664	90-89-88
16	Técnica Aseguradora	C-352	90-89-88-87-86
17	Unión peninsular de seguros	C-555	90-89-88-87-86
18	Mas Grupo 86 Fondo Asegurador	C-581	89-88-87-86-85
19	Munauto <b>(4)</b>	C-608	89-88-87
20	Unión Ibérica Grupo 86	C-523	89-88-87-86-85
21	Sociedad occidental de Seguros <b>(5)</b>	C-615	89-88
22	España Vitalicia	C-071	89-88-87-86-85
23	Instituto Médico Quirúrgico	C-422	88-87-86-85-84
24	Madrid, S.A. de seguros Generales	C-111	86-85-84-83
25	Cía. Mercantil de Seguros	C-560	86-85-84-83
26	Médica Riojana	C-460	86-85-84-83
27	Igualatorio Vallisoletano Médico Quirur. y de Espec.	C-328	86-85
28	Igualatorio Médico Ntra. Sra. Del Rosario	C-321	86-85-84
29	Argüelles	C-289	86-85-84-83
30	Palace	C-250	85-84-83
31	Asociación clínica Española	C-283	84
32	Cosmos	C-564	84-83
33	Alianza Previsora	C-220	84-83
34	Clínica Argüeso	C-224	84-83
35	Labor	C-330	84-83
36	La Gloria Eterna	C-474	84-83

**Notas**

- (1)- Asistencia Sanitaria 2000, antes Policlínica Santiago
- (2) - Mundi Seguros empezó a operar en noviembre del 88
- (3)- Mades Fondo Asegurador, antes mutua.
- (4)- Munauto, antes Mutua
- (5)- Sociedad Occidental de Seguros, solo operó dos años



**TABLA 4.2. EMPRESAS SANAS**

Nº	NOMBRE	CODIGO	AÑOS
101	Metrópolis	C-121	93-92-91-90-89
102	Igualatorio Médico Leonés	C-403	93-92-91-90-89
103	Alergia	C-286	93-92-91-90-89
104	Seguros Mercurio	C-630	93-92-91-90-89
105	Génesis Seguros Generales <b>(1)</b>	C-695	93-92-91-90
106	Lagún-Aro	C-572	92-91-90-89-88
107	La Unión Alcoyana	C-188	92-91-90-89-88
108	Athena, cía. Ibérica de seguros y reaseg. <b>(2)</b>	C-228	91-90-89-88-87
109	Lepanto <b>(3)</b>	C-108	91-90-89-88-87
110	Federación Ibérica <b>(4)</b>	C-076	91-90-89-88-87
111	La patria Hispana	C-139	90-89-88-87-86
112	Asociación Médica conquense	C-313	90-89-88-87-86
113	Aseguradora Universal	C-012	90-89-88-87-86
114	Sur	C-186	90-89-88-87-86
115	Munat <b>(5)</b>	C-665	90-89-88
116	Europa Seguros Diversos <b>(6)</b>	C-508	90-89-88-87-86
117	Hispano Alsaciana <b>(7)</b>	C-061	90-89-88-87-86
118	Nortehispania	C-275	89-88-87-86-85
119	Andalucía y Fenix agrícola <b>(8)</b>	C-004	89-88-87-86-85
120	Compañía Astra	C-468	89-88-87-86-85
121	La Alianza española	C-002	89-88-87-86-85
122	La humanitaria <b>(9)</b>	C-318	89-88-87-86-85
123	Boreal Médica	C-027	88-87-86-85-84
124	ADEA <b>(10)</b>	C-378	86-85-84-83
125	ASEFA	C-522	86-85-84-83
126	Igualatorio Médico Quirurg. Pilarista <b>(11)</b>	C-390	86-85-84-83
127	Asistencia Clínica Univ. De Navarra	C-325	86-85-84
128	Sanitaria Médico Quirúrgico	C-515	86-85-84-83
129	La Antártida	C-506	86-85-84-83
130	La Gloria	C-229	85-84-83
131	Federación Médica	C-434	84-83
132	Le Mans S.E.	C-552	84
133	Clinos	C-226	84-83
134	Salus, Asistencia Sanitaria	C-485	84-83
135	El Paraíso Universal	C-238	84-83
136	Seguro Europeo	C-319	84-83

## Notas

- (1)- Génesis, año 91 primer balance completo. Cenit sustituye a Génesis
- (2)- Athena sustituye a DAPA
- (3)- Lepanto sustituye a Regal Insurance
- (4)- Federación Ibérica sustituye a AMIC
- (5)- Munat, antes Mutua
- (6)- Europa Seguros Diversos, antes Previsur
- (7)- Hispano Alsaciana, absorbe Azur y cambia de denominación
- (8)- Andalucía y Fenix Agrícola, actual Hispasalud
- (9)- La Humanitaria, actual Al-Andalus
- (10)- ADEA, transformación en Gan-España, C-686
- (11)- Igualatorio M. Q. Pilarista, actual Pilarista

### **II.3.- Análisis del balance y la cuenta de resultados de las entidades aseguradoras**

Tras la obtención de la muestra de empresas hemos procedido a la homogeneización de la información contable para tener una reclasificación operativa de los estados contables.

Es necesario mencionar que, derivado del periodo que comprende la muestra, tanto el balance como la cuenta de resultados que vamos a reclasificar se corresponden con el modelo vigente hasta 1997 (Plan General de Contabilidad adaptado a las entidades de seguros, reaseguros y capitalización aprobado por la O.M. de 30 de julio de 1981). En la actualidad el Plan de Cuentas que contiene los modelos de Cuentas Anuales que están en vigor en la actualidad es el aprobado por el Real Decreto 2014/1997.

También queremos resaltar que dadas las peculiaridades que presenta el sector asegurador, existen ciertas dificultades conceptuales a la hora de asignar determinadas partidas al circulante o al fijo (por ejemplo las provisiones técnicas o las inversiones). Este es el motivo por el cual, aunque en nuestro trabajo hemos seguido el criterio tradicional de partir de la división del activo y pasivo de una sociedad aseguradora según el balance propuesto por el plan sectorial de contabilidad y ordenar el balance en distintas masas patrimoniales con el criterio de menor a mayor liquidez en el activo y de menor a mayor exigibilidad en el pasivo, tal y como expondremos a continuación, caben otros criterios a la hora de clasificar las partidas entre fijo y circulante que afectarían a la definición de las masas patrimoniales y, en consecuencia a la definición de los ratios (Linares, 1998, p. 45-50).

La utilización de otras clasificaciones (Millán, 2000, p.82-85; Fernández-Palacios y Maestro, 1991, p. 518-527; García y Ballester, 1994, p. 162-163) dependerá del usuario del modelo y afectará a los resultados del modelo y a

su interpretación, pero el objeto de este trabajo es mostrar la utilidad de la utilización de métodos multicriterio en la predicción de las crisis empresariales con independencia del sector que se considere y de los criterios utilizados en la elaboración de los datos de entrada de los modelos, ya sean estos contables o no.

A continuación se expone el balance y la cuenta de pérdidas y ganancias para el sector asegurador con los códigos asignados a las partidas que van a utilizarse para el cálculo de los ratios, así como la estructura de reclasificación para el balance:

**BALANCE DE SITUACIÓN**

<b>I.- ACCIONISTAS por desembolsos no exigidos</b>	<b>A1</b>	<b>I.- CAPITALES PROPIOS</b>	<b>P1</b>
<b>II.- INMOVILIZADO</b>	<b>A2</b>	1.- Capital suscrito, Fondo Mutual o Fondo Permanente	P11
1.- Gastos de Establecimiento y otros amortizables	A21	2.- Primas de emisión	P12
2.- Inmaterial	A22	3.- Diferencias por actualizaciones del activo	P13
3.- Material	A23	4.- Reservas	P14
<b>III.- INVERSIONES</b>	<b>A3</b>	5.- Resultados de ejercicios anteriores pendientes de aplic.	P15
1.- Materiales	A31	6.- Resultado del ejercicio después del impuesto	P16
2.- Financieras	A32	7.- Minusvalías en valores negociables de renta fija	P17
3.- Inversiones en empresas del grupo, asoc. y particip. y acciones ppas.	A33	<b>II.- PROVISIONES TÉCNICAS</b>	<b>P2</b>
<b>IV.- PROVISIONES TÉCNICAS DEL REASEGURO CEDIDO Y RETROC.</b>	<b>A4</b>	1.- Provisiones técnicas para riesgos en curso	P21
1.- Provisiones técnicas para riesgos en curso	A41	2.- Provisiones matemáticas (vida)	P22
2.- Provisiones matemáticas (vida)	A42	3.- Provisiones técnicas para prestaciones	P23
3.- Provisiones técnicas para prestaciones	A43	4.- Otras provisiones técnicas	P24
4.- Otras provisiones técnicas	A44	<b>III.- PROVISIONES PARA RESPONSABILIDADES Y GASTOS</b>	<b>P3</b>
<b>V.- CRÉDITOS</b>	<b>A5</b>	<b>IV.- DEPÓSITOS RECIBIDOS POR REASEG. CEDIDO Y RETROC.</b>	<b>P4</b>
1.- Entidades y Pools de Seguros y Reaseguros	A51	<b>V.- DEUDAS</b>	<b>P5</b>
2.- Créditos contra agentes	A52	1.- Empréstitos	P51
3.- Provisiones (a deducir)	A53	2.- Deudas a establecimientos de crédito	P52
4.- Créditos contra asegurados	A54	3.- Entidades y Pools de Seguros y Reaseguros	P53
5.- Créditos fiscales, sociales y otros	A55	4.- Deudas con agentes	P54
6.- Accionistas por los desembolsos exigidos	A56	5.- Deudas con asegurados	P55
7.- Dividendos activos a cuenta	A57	6.- Deudas condicionadas	P56
8.- Provisiones (a deducir)	A58	7.- Deudas a empresas del grupo	P57
<b>VI.- CUENTA DE AJUSTE POR PERIODIFICACIÓN</b>	<b>A6</b>	8.- Deudas a empresas asociadas y participadas	P58
1.- Gastos anticipados e intereses pagados por anticipado	A61	9.- Operaciones preparatorias o complementarias de seguros de vida, no acogidas a la Ley 8/1987 de 8 de junio.	P59
2.- Otras cuentas de periodificación	A62	10.- Cuentas fiscales, sociales y otras	P51
<b>VII.- EFECTIVO EN BANCOS Y OTROS ESTABLECIMIENTOS DE CRÉDITO, EN CAJA Y EN CHEQUES</b>	<b>A7</b>	<b>VI.- CUENTAS DE AJUSTE POR PERIODIFICACIÓN</b>	<b>P6</b>
<b>TOTAL ACTIVO I+II+III+IV+V+VI+VII</b>	<b>AA</b>	<b>TOTAL PASIVO I+II+III+IV+V+VI</b>	<b>PP</b>

## CUENTA DE PÉRDIDAS Y GANANCIAS

	SEGURO DIRECTO	REASEG. CEDIDO MÁS RETROC. (-)	NEGOCIO NETO		SEGURO DIRECTO	REASEG. CEDIDO MÁS RETROC. (-)	NEGOCIO NETO
<b>I.- GASTOS TÉCNICOS</b>							
1.- Gastos técnicos no vida:							
1.1.- Prestaciones y gastos pagados, no vida	DD1111	DR1111	D1111		HD1111	HR1111	H1111
1.2.- Provisiones técnicas para prestaciones, no vida:							
+Al cierre del ejercicio	DD11121	DR11121	D11121		HD11121	HR11121	H11121
-Al comienzo del ejercicio	DD11122	DR11122	D11122		HD11122	HR11122	H11122
1.3.- Otras provisiones técnicas, no vida							
+ Al cierre del ejercicio	DD11131	DR11131	D11131		HD11131	HR11131	H11131
- al comienzo del ejercicio	DD11132	DR11132	D11132		HD11132	HR11132	H11132
2.- Gastos técnicos vida:							
2.1.- Prestaciones y gastos pagados, vida	DD1221	DR1221	D1221		HD1221	HR1221	H1221
2.2.- Provisiones técnicas para prestaciones, vida:							
+Al cierre del ejercicio	DD12221	DR12221	D12221		HD12221	HR12221	H12221
-Al comienzo del ejercicio	DD12222	DR12222	D12222		HD12222	HR12222	H12222
2.3.- Provisiones matemáticas, vida:							
+Al cierre del ejercicio	DD12231	DR12231	D12231		HD1	HR1	H1
-Al comienzo del ejercicio	DD12232	DR12232	D12232				
2.4.- Otras provisiones técnicas, vida:							
+Al cierre del ejercicio	DD12241	DR12241	D12241				
-Al comienzo del ejercicio	DD12242	DR12242	D12242				
<b>TOTAL GASTOS TÉCNICOS VIDA Y NO VIDA</b>	DD1	DR1	D1				
<b>II.- COMISIONES Y OTROS GASTOS DE EXPLOTACIÓN</b>							
1.- Comisiones y participaciones:							
1.1.- Comisiones no vida del ejercicio	DD2111		D2111				
1.2.- Comisiones vida del ejercicio							
+ Comisiones y participac. de primas devengadas del año	DD21121		D21121				
- Comisiones del año llevadas al activo	DD21122		D21122				
+ Amortización en el año de las comisiones de adquisición llevadas al activo.	DD21123		D21123				
1.3.- Gastos de agencia	DD2113		D2113				
2.- Otros gastos de explotación							
2.1.- Sueldos y salarios			D2221				
2.2.- Cargas sociales			D2222				
2.3.- Dotaciones del ejercicio para amortizaciones			D2223				
2.4.- Dotaciones a las provisiones			D2224				
2.5.- Gastos de explotación varios			D2225				
3.- Comisiones y participaciones del reaseguro							
No vida			D231				
<b>I.- PRIMAS Y RECARGOS</b>							
1.- Primas adquiridas, no vida							
1.1.- Primas y recargos netos de anulaciones, no vida	HD1111	HR1111			HD1111	HR1111	H1111
1.2.- Provisiones técnicas para riesgos en curso, no vida:							
+ Al comienzo del ejercicio	HD11121	HR11121			HD11121	HR11121	H11121
- Al cierre del ejercicio	HD11122	HR11122			HD11122	HR11122	H11122
1.3.- Provisiones para primas pendientes, no vida:							
+ Al comienzo del ejercicio	HD11131	HR11131			HD11131	HR11131	H11131
- Al cierre del ejercicio	HD11132	HR11132			HD11132	HR11132	H11132
2.- Primas adquiridas, vida:							
2.1.- Primas y recargos netos de anualaciones, vida	HD1221	HR1221			HD1221	HR1221	H1221
2.2.- Provisiones para primas pendientes, vida:							
+ Al comienzo del ejercicio	HD12221	HR12221			HD12221	HR12221	H12221
- Al cierre del ejercicio	HD12222	HR12222			HD12222	HR12222	H12222
<b>TOTAL PRIMAS ADQUIRIDAS VIDA Y NO VIDA</b>	HD1	HR1			HD1	HR1	H1
<b>II.- OTROS INGRESOS DE EXPLOTACIÓN</b>							
1.- Ingresos accesorios a la explotación							H21
2.- Provisiones aplicadas a su finalidad							H22
<b>TOTAL OTROS INGRESOS DE EXPLOTACIÓN</b>							H2
<b>III.- INGRESOS FINANCIEROS</b>							
1.- Ingresos de inversiones materiales							H31
2.- Ingresos de inversiones financieras							H32
3.- Ingresos de invers. en empresas del grupo, asoc. y particip.							H33
4.- Ingresos financieros varios							H34
5.- Provisiones aplicadas a su finalidad							H35
6.- Beneficios por diferencias de cambio en divisas							H36
7.- Beneficios en realización de inversiones materiales							H37
8.- Beneficios en realización de inversiones financieras							H38
<b>TOTAL INGRESOS FINANCIEROS</b>							H3
<b>IV.- BENEFICIOS EXCEPCIONALES</b>							H4
<b>TOTAL I+II+III+IV</b>							
<b>V.- PÉRDIDAS DEL EJERCICIO (Saldo que pasa al balance)</b>							H5
<b>TOTAL GENERAL</b>							HH

Vida	D232
<b>TOTAL COMISIONES Y OTROS GASTOS DE EXPLOTACIÓN</b>	D2
<b>III.- GASTOS FINANCIEROS</b>	
1.- Gastos de inversiones materiales (incluidas amortizaciones)	D31
2.- Gastos de inversiones financieras	D32
3.- Gastos inversiones en empresas del grupo, asociadas y participadas	D33
4.- Gastos financieros varios	D34
5.- Dotación del ejercicio para provisiones	D35
6.- Pérdidas por diferencias de cambio de divisas	D36
7.- Pérdidas en realización de inversiones materiales	D37
8.- Pérdidas en realización de inversiones financieras	D38
<b>TOTAL GASTOS FINANCIEROS</b>	D3
<b>IV.- PÉRDIDAS EXCEPCIONALES</b>	D4
<b>TOTAL I+II+III+IV</b>	
<b>V.- IMPUESTO SOBRE SOCIEDADES</b>	D5
<b>VI.- BENEFICIO DEL EJERCICIO DESPUÉS DEL IMPUESTO</b> (Saldo que pasa al balance)	D6
<b>TOTAL GENERAL</b>	DD

*Reclasificación del Balance*

**Activo**

<p>I. Accionistas                  II. Inmovilizado                  III. Inversiones                      1. Materiales                      2. Financieras                      3. Empresas del grupo</p>	<p><b>Activo Fijo</b>                  I. Accionistas                  II. Inmovilizado                  III. Inversiones                      1. Materiales                      2. Financieras <b>(1)</b>                      3. Empresas del grupo</p>
<p>IV. Provisiones de reaseguro                  V. Créditos</p>	<p><b>Circulante exigible</b>                  IV. Provisiones de reaseguro                  V. Créditos</p>
<p>VI. Ajustes</p>	<p><b>Circulante realizable</b>                  III. Inversiones                      2. Financieras <b>(2)</b>                  VI. Ajustes</p>
<p>VII. Efectivo</p>	<p><b>Circulante Disponible</b>                  VII. Efectivo</p>

**(1)** Dentro de las inversiones financieras solo las de control (cobertura de provisiones técnicas) no considerando las especulativas. Si las inversiones en empresas del grupo cotizasen en el mercado organizado pasarían a circulante exigible.

**(2)** Solo las que coticen en mercados organizados.

*Nota:* A los efectos del cálculo de los ratios y debido a que no tenemos el desglose de las inversiones (únicamente disponemos del dato que figura en el balance) las incluiremos siempre como parte del activo



circulante, es decir como si no existiese cartera de control o que todas las inversiones cotizaran en mercados organizados.

**Pasivo**

I. Capitales propios	<b>Neto Patrimonial</b> I. Capitales propios
II. Provisiones técnicas 1. Riesgos en curso 2. Matemáticas 3. Para prestaciones 4. Otras III. Provisión de responsabilidades IV: Depósitos de reaseguros V. Deudas	<b>Exigible a largo plazo</b> II. Provisiones técnicas 1. Riesgos en curso 2. Matemáticas 4. Otras III. Provisión de responsabilidades IV: Depósitos de reaseguros V. Deudas <b>(1)</b>
VI. Ajustes	<b>Exigibles a corto plazo</b> II. Provisiones técnicas 3. Para prestaciones V. Deudas a corto plazo VI. Ajustes

**(1) Deudas a largo plazo**

*Nota:* A los efectos del cálculo de los ratios y debido a que no tenemos el desglose de las deudas (únicamente disponemos del dato que figura en el balance) consideraremos que son a largo plazo los empréstitos, las deudas con empresas del grupo y las deudas con empresas asociadas. Es necesario mencionar que las deudas y establecimientos de crédito (epígrafe V.2.) han sido incluidas en el corto plazo pero podrían haberse considerado a largo plazo. La arbitrariedad viene

derivada de, como en el caso anterior, la falta de desglose de las deudas.

Una vez estructurado el balance de la sociedad, se pueden efectuar un análisis financiero, comparativo (con una empresa modelo, una empresa ideal, un balance agregado sectorial) o calcular ratios.

## II.4.- Elección De Las Variables: Los Ratios Contables

La elección de las variables más adecuadas a utilizar en la elaboración del modelo de predicción es parte fundamental del éxito final del mismo (García *et al.*, 1997). En este trabajo utilizaremos fundamentalmente ratios financieros, ya que el análisis financiero es uno de los sistemas más empleados en predicción de insolvencias.

Además, al tratarse de un modelo de predicción del fracaso empresarial se deben considerar posibles variables significativas que discriminen la insolvencia. De esta forma, hemos configurado un conjunto de variables, que, a nuestro juicio, y, desde un punto de vista teórico y práctico, forman un modelo integral de análisis financiero de una entidad aseguradora, y que consideramos que son las más relevantes.

Ya anticipamos, pero ahora lo concretamos, que en nuestro estudio no vamos a partir de una gran batería de ratios. Introducir muchas variables en un modelo puede acarrear el efecto contrario al que buscamos. Es decir, si metemos muchas variables en un modelo, en vez de incrementar la capacidad predictiva, estamos introduciendo “ruido” en el mismo que puede disminuir dicha capacidad.

Además, en el epígrafe anterior hemos justificado por qué utilizamos ratios financieros, que son variables de tipo cuantitativo. Aunque la metodología rough set permite el uso de variables cualitativas y, pensamos que puede ser muy positivo su introducción en cualquier modelo, no hemos considerado características cualitativas debido a la imposibilidad de conseguirlas.

En cuanto a la selección de los ratios, hemos acudido a los más utilizados en el análisis financiero para el pronóstico de la insolvencia de entidades aseguradoras, teniendo en cuenta nuestro criterio personal, dadas

las propias características del sector seguros y teniendo en cuenta un análisis preliminar de las particularidades de los dos grupos de empresas (sanas y fracasadas).

Es necesario recalcar que las peculiaridades sectoriales de las empresas de seguros, son tan significativas, que no resultan aplicables, al menos parcialmente, los modelos y contenidos de las Cuentas anuales elaboradas para empresas industriales y comerciales (Prieto, 1992, p.15), por lo que junto con los ratios aplicables al conjunto de empresas industriales y comerciales hay que considerar todos aquellos específicos para el sector asegurador.

Pretendemos, posteriormente con la metodología Rough Set, obtener el más pequeño subconjunto de ratios con el mayor poder discriminante en relación con el fracaso empresarial.

La inversión del proceso productivo que caracteriza al sector asegurador implica la determinación a priori del precio del seguro o de la prima en base a los valores esperados tanto de la siniestralidad como de los gastos. Al hablar de valores esperados, cabe la posibilidad de que se produzcan desviaciones negativas entre los datos reales y los previstos. Por tanto la empresa debe disponer de unos fondos patrimoniales o margen de solvencia para hacer frente a las posibles desviaciones.

El margen de solvencia se convierte, de este modo, en una variable clave en cualquier estudio que implique la viabilidad de las empresas del sector asegurador. En nuestra propuesta de ratios existían varios que incluían dicha variable en su definición pero debido al período que abarca la muestra dicho dato se encuentra en los estados confidenciales remitidos a la Dirección General de Seguros que no son públicos, y, en consecuencia, no se puede tener acceso a ellos con fines de investigación. Este es el motivo por el que en ninguno de nuestros ratios aparece el margen de solvencia

pero creemos es una variable clave que en muestras que abarquen periodos posteriores en los que dicho dato ya es público debería aparecer en al menos uno de los ratios.

Vamos a describir brevemente dichos ratios y justificar su elección para formar parte de los modelos.

Es necesario resaltar, que a lo largo de la literatura contable se han propuesto muchos ratios, tanto generales como específicos para el sector asegurador, y por tanto, otros podrían haberse utilizado<sup>1</sup>, pero por los motivos que mencionaremos estos son los que hemos aplicado para medir la solvencia financiera de las entidades aseguradoras. Aquí queremos resaltar la misma advertencia que efectuamos al realizar la reclasificación del balance, el objeto de este trabajo es mostrar la utilidad de la utilización de métodos multicriterio en la predicción de las crisis empresariales, la flexibilidad del método utilizado permite que las variables puedan ser cualitativas o cuantitativas, o que los ratios sean unos u otros o cambiarlos en función de las condiciones económicas o del sector que se considere, o que el número de ellos sea mayor o menor.

A continuación pasamos a describir los ratios.

---

<sup>1</sup> Hay ratios financieros convencionales que se utilizan en muchos estudios sobre solvencia empresarial y que no aparecen en nuestro trabajo. Por citar alguno de ello estaría el ratio Activo circulante/Pasivo Circulante o Activo Fijo/Deudas a largo plazo pero por las razones que indican García y Ballester (1994), carecen de utilidad operativa para un correcto análisis de la solvencia empresarial por basarse en una concepción patrimonialista de la empresa, es decir sin tener en cuenta el principio de gestión continuada y comparando partidas, si bien homogéneas en cuanto al periodo medio de maduración, son heterogéneas para el análisis financiero.

Ratio del capital circulante o fondo de maniobra sobre total activo

$$R1 = \frac{\text{Fondo de Maniobra}}{\text{Total Activo}}$$

En ocasiones recibe la denominación de liquidez la relación existente entre el capital circulante y el activo total de la empresa. El objetivo de este ratio es medir la liquidez neta de los activos sobre el total de la capitalización, significando la importancia que representa el capital circulante en relación con el total activo y, por tanto, la garantía que ofrece la empresa con su activo circulante para afrontar a un pasivo de igual grado (Álvarez y Urías, 1987).

Cuanto mayor sea el valor de este ratio indicará una mejor situación respecto a la liquidez. El problema de liquidez es bastante importante en una empresa aseguradora ya que dada la inversión del proceso productivo los cobros se realizan antes que los pagos luego no debería existir este problema. Además si una empresa aseguradora no es capaz de pagar los siniestros que se producen puede acarrear una pérdida de la confianza de sus clientes.

Además, la empresa debe ofrecer un equilibrio entre su estructura económica y financiera. Para obtener este equilibrio las fuentes financieras deben ajustarse, de acuerdo a un plan financiero, a las inversiones. Los elementos acíclicos y circulantes necesarios para la continuidad del ciclo de explotación tienen que estar financiados con recursos propios y recursos ajenos a largo plazo en su correcta proporción. El desequilibrio financiero más frecuente, surge, generalmente, del inadecuado diseño de la financiación de los inmovilizados, lo que puede originar a corto plazo serios problemas de liquidez o tensiones financieras innecesarias (García *et al.*, 1997, p. 141).

El equilibrio financiero lo mediremos a través del cálculo del fondo de maniobra (diferencia entre los activos circulantes y pasivos circulantes), y para facilitar la comparación determinamos el porcentaje que representa sobre el total del activo.

Cuando el fondo de maniobra es positivo, significa que parte del activo circulante está siendo financiado con recursos permanentes, ya sean propios o ajenos a largo plazo. Por el contrario, un capital circulante negativo, implica que una parte del inmovilizado está siendo financiado por deudas a largo plazo.

*Ratio de rentabilidad financiera de los recursos propios*

$$R2 = \frac{\text{Beneficio antes de Impuestos}}{\text{Capitales Propios}}$$

Con carácter general, la rentabilidad financiera de una empresa trata de comparar el rendimiento obtenido, a lo largo del ejercicio económico considerado, con los capitales propios invertidos (Rivero, 1987, p. 239-240).

El marco genérico en el que se envuelve toda empresa gira en torno a dos premisas fundamentales, asegurar su funcionamiento en el futuro bajo hipótesis de normalidad, y conseguir que ésta se desarrolle en las mejores condiciones de rentabilidad posibles. Para la empresa, el objetivo del beneficio debe cumplir una doble función, por un lado, retribuir convenientemente a los suministradores del capital, y por otro capitalizar la empresa de forma que se asegure un correcto equilibrio financiero (García *et al.*, 1997, p. 146).

En consecuencia, para una empresa es necesario conseguir una adecuada rentabilidad financiera no solo por razones de imagen que le

permitan colocar las ampliaciones de capital o conseguir créditos sino también porque un beneficio suficiente en una firma con autonomía financiera lleva consigo el mantenimiento de la misma.

Este ratio informa sobre la eficiencia de la empresa en la utilización de los capitales invertidos por accionistas (o de los acreedores ya que también se considera como rentabilidad financiera al ratio que relaciona el beneficio con el pasivo).

Si la rentabilidad financiera es mayor que cero, se dice que existe una rentabilidad financiera *positiva* en relación con los recursos propios (o con el pasivo, según se considere).

Además, de acuerdo con el estudio realizado por Gabás (1997, p. 26), en el que se analizan distintos trabajos empíricos sobre predicción de la insolvencia empresarial, la variable más repetida en cuanto a eficiencia en su capacidad de predicción es la rentabilidad o capacidad de obtener beneficios por lo que habría que destacar su importancia desde el punto de vista de aportación a una teoría general de la solvencia.

*Ratio de ingresos financieros por activo de inversiones*

$$R3 = \frac{\text{Ingresos Financieros}}{\text{Activo de Inversiones}}$$

Es un ratio que se considera que puede modificar la posición de solvencia ya que informa sobre la eficiencia en el empleo e inversión de los recursos financieros y la proporción que estos suponen respecto a activos. Por tanto expresa la eficiencia en el aprovechamiento de los activos dedicados a inversiones, y en consecuencia, la medida de la bondad en la política de inversiones de los recursos disponibles (Martín *et al.*, 1999, p. 80).



Este ratio cobra aún más importancia si tenemos en cuenta que los rendimientos financieros de las entidades aseguradoras son en la actualidad la fuente principal de sus beneficios, o lo que es lo mismo, representan la posibilidad de obtener una rentabilidad adecuada para los recursos propios y también una fuente de financiación importante de los incrementos del Margen de Solvencia (Prieto, 1993, p. 95).

Además, es uno de los ratios utilizado por la NAIC (National Association Insurance Commissioners) en su sistema IRIS (Insurance Regulatory Information System), sistema de alarma para la detección de posibles insolvencias. Nos indica una medida de la adecuación del rendimiento obtenido por la cartera de inversiones.

*Ratio de cash flow sobre total pasivo*

$$R4 = \frac{\text{Cash Flow}}{\text{Total Pasivo}}$$

La justificación de utilizar el cash-flow (como recursos generados) viene motivada por la capacidad que tiene las empresas para flexibilizar, alisar e incluso manipular el resultado contable. Por ello, se ha desarrollado una línea de investigación para contrastar si los fondos de tesorería generados son más eficaces para la predicción de la solvencia, ya que son menos manipulables si los comparamos con los resultados. En *casos específicos*, la evolución de los flujos de tesorería es *mucho más explicativa* de los problemas reales de la empresa que el beneficio (Gabás, 1997, p.27).

La caída final que supone un desplome de beneficios a cuantiosas pérdidas, que hemos podido comprobar en algunas cuentas anuales nos hace sospechar de una política de manipulación del mismo en los últimos años de la vida de las empresas, aunque tanto en nuestro estudio como en otros trabajos de investigación, partimos de la hipótesis de adecuada calidad

de la información contable, por lo que si esto no sucede se produce un sesgo imposible de controlar.

Además según se constata de los informes de auditoría que acompañan las cuentas anuales de las empresas en suspensión de pagos, en un alto porcentaje significativo de casos, se dan incumplimientos de principios contables (Roqueta, 2000, p. 2-5). Ante las dificultades económicas, la empresa disminuye o dota insuficientemente las cuentas de amortizaciones y provisiones y suelen realizar operaciones extraordinarias. Estas últimas son las que han motivado que hayamos considerado el cashflow añadiendo los recursos generados por los resultados extraordinarios.

En el trabajo realizado (ver referencia anterior), este ratio detecta precozmente la insolvencia (al situarla por debajo del promedio del mercado, representado por el conjunto de todas las empresas) motivo por el cual hemos considerado incluirlo en nuestro estudio. Además podemos mencionar otros trabajos (Laitinen, 1992, p. 323-340; Karels y Prakash, 1987, p. 573-592), en los cuales este ratio se ha mostrado como el más eficiente a la hora de predecir el fracaso empresarial. El argumento que justifica esta eficiencia en cuanto a su capacidad de predicción es, que una financiación escasa debida a ingresos insuficientes que aparezcan de manera inesperada puede ser la causa del comienzo de proceso de quiebra. Los motivos que la pueden originar son varios: baja rentabilidad, crecimiento demasiado rápido o o una velocidad del capital demasiado lenta (entendiendo por esta última la capacidad de los activos para generar ingresos).

Ratios de solvencia en sentido estricto

Según hemos visto, la solvencia de una sociedad es un concepto difícil de definir con precisión. El mercado asegurador al tratarse de un mercado institucional y depender de la confianza del consumidor en la institución y de la eficacia del marco jurídico en el que se desenvuelve, hace que la noción de solvencia en la entidad aseguradora tenga matices propios (Linares, 2000, p.59). De una parte es función de los activos realizables y, de otra, de las obligaciones ciertas que hubiera asumido, pero también de conceptos como la rentabilidad de la cartera o desarrollo futuro de la misma.

En las entidades aseguradoras distinguimos dos tipos de solvencia:

*Solvencia estática*, que indica la disponibilidad del asegurador de medios suficientes para hacer frente a sus obligaciones contractuales. Este concepto podría, en términos generales, identificarse con el volumen de provisiones técnicas y su correcta inversión.

*Solvencia dinámica*, que se basa en medir la capacidad de la sociedad para hacer frente, no ya a sus obligaciones contractuales, sino a las futuras que pudiera adquirir.

Se refiere a la capacidad de la empresa para seguir desempeñando en el futuro la función de cobertura de riesgos que le es propia (Fernández-Palacios y Maestro, 1991).

Al analizar una entidad aseguradora debemos considerar la posibilidad de que se produzca una siniestralidad superior a la técnicamente prevista. Es decir, dado que la prima representa el valor medio de la siniestralidad esperada y es una variable aleatoria, ésta puede tomar infinitos valores en torno a ese valor medio que den lugar a que, por mucho que esté bien calculada, la siniestralidad real no coincida con la esperada, sino que en

unos ejercicios se halle por debajo y en otros por encima. En este último caso, si la entidad no contara con más recursos que las primas recaudadas, la entidad entraría en insolvencia porque con ellas no podría hacer frente a todos los siniestros (Maestro, 2000).

En consecuencia con lo expuesto en el párrafo precedente, surge además otra necesidad: la de tener un patrimonio o fondo de cuantía suficiente, cuya finalidad sea la de compensar la insuficiencia de los ingresos percibidos por la entidad en relación con los gastos, cualquiera que sea la causa que los provoque (Fernández-Palacios y Maestro, 1991).

Plantear estas posibilidades, es plantear el problema de la solvencia dinámica de una entidad aseguradora.

Por tanto se hace necesario que la entidad disponga de un patrimonio complementario, es decir, un patrimonio no afecto a obligación alguna. Estas garantías financieras adicionales que persiguen que el asegurador cumpla el principio de solvencia, en su acepción dinámica, son el margen de solvencia y la provisión de estabilización.

$$R5 = \frac{\text{Total Primas adquiridas de seguro directo}}{\text{Capitales Propios}}$$

$$R6 = \frac{\text{Total Primas adquiridas de negocio neto}}{\text{Capitales Propios}}$$

$$R7 = \frac{\text{Total Primas adquiridas de seguro directo}}{\text{Capitales Propios} + \text{Provisiones Técnicas}}$$

$$R8 = \frac{\text{Total Primas adquiridas de negocio neto}}{\text{Capitales Propios} + \text{Provisiones Técnicas}}$$

Estos cuatro ratios (al igual que **R11-R12-R13-R14** que comentaremos a continuación) se pueden considerar como indicadores de solvencia en un

momento determinado del tiempo, solvencia estática, y también, estudiando su variación a lo largo de un determinado periodo, serían indicadores de la solvencia dinámica (Martín *et al.*, 1999, p. 77-78).

Estos ratios recogen en sus numeradores la medida de los riesgos anuales, basándose en la cuantía de las primas destinadas a cubrir los riesgos asegurados en el año (primas devengadas en el período en que se trate). Se diferencia entre seguro directo y seguro neto, con el objetivo de comprobar las diferencias entre ambas posiciones y observar su repercusión en la solvencia.

Los denominadores muestran el soporte financiero de las empresas. En los dos primeros se contemplan los capitales propios en sentido estricto lo que sugiere un soporte global del riesgo independiente de las características anuales. En los dos últimos se suman a los capitales propios las provisiones técnicas de las que se dispone (totales o netas según el tipo de gasto técnico con el que se compare), logrando de este modo el soporte financiero real para el periodo analizado.

Estos ratios, que utilizan las primas como una medida del riesgo en que incurre la empresa de seguros, son usados en Estados Unidos por la *National Association of Insurance Comisioners* (N.A.I.C.), dentro de sus ratios de alarma ante una posible situación de insolvencia.

Además los dos primeros ratios (el segundo es el primero expresado neto de reaseguro cedido) se denominan *Ratios de Apalancamiento Asegurado* (Millán, 2000, p. 103-104). Reflejan el apalancamiento del volumen de negocio neto de la compañía en relación con sus recursos propios. Mide el grado de absorción por la compañía de los errores en el precio de sus productos. Estos dos ratios no es aconsejable que presenten una cuantía muy elevada, ya que el nivel de apalancamiento incrementa el retorno del capital (con pocos recursos se está generando un elevado

volumen de negocios, *rotación de los recursos permanentes* (Linares, 2000, p. 55) pero también el riesgo de inestabilidad.

$$R11 = \frac{\text{Gastos técnicos de seguro directo}}{\text{Capitales Propios}}$$

$$R12 = \frac{\text{Gastos técnicos de negocio neto}}{\text{Capitales Propios}}$$

$$R13 = \frac{\text{Gastos técnicos de seguro directo}}{\text{Capitales Propios} + \text{Provisiones Técnicas}}$$

$$R14 = \frac{\text{Gastos técnicos de negocio neto}}{\text{Capitales Propios} + \text{Provisiones Técnicas}}$$

Estos cuatro ratios, **R11-R12-R13-R14**, se diferencian de los anteriores en que recogen en sus numeradores la medida de los riesgos anuales, basándose en la valoración de los riesgos que realmente han ocurrido (siniestros del año) que se recogen en la Cuenta de Pérdidas y Ganancias como Gastos Técnicos. Al igual que en los anteriores, se diferencia entre seguro directo y neto para poder analizar la influencia del reaseguro en la estructura de solvencia.

*Ratio de autonomía financiera*

$$R9 = \frac{\text{Capitales Propios}}{\text{Pasivo Total}}$$

El objetivo de este ratio es informar sobre la composición estructural de las fuentes de financiación y medir el grado de autonomía o independencia financiera de la entidad (en el sentido de si ante una decisión de inversión puede elegir libremente los recursos que más le interesen) en virtud de la procedencia de las fuentes financieras utilizadas (Rivero *et al.*, 1998, p. 202-203).

La autonomía financiera implica la posibilidad de elegir, por parte de la empresa, entre las distintas fuentes de financiación, las que considera más adecuadas para acometer la explotación de un proceso productivo. En el caso de no poder tomar los fondos que le resultaran más convenientes por tener cerradas otras fuentes de financiación, la empresa habrá perdido su independencia y no tendrá libertad financiera que garantice la autonomía de sus decisiones. La firma, en este caso, actuará de una manera forzada.

A través de este ratio se intenta conocer cuál es el nivel óptimo (proporción) de endeudamiento de una empresa, o sea, cuál es el grado de independencia con respecto al exterior, con terceros. Evalúa la cantidad de recursos que son propiedad de los accionistas. Cuanto mayor sea el ratio que analizamos menor será la probabilidad de insolvencia a largo plazo, así como menores serán las necesidades de la empresa para hacer frente al pago de la carga financiera y al reembolso del principal. En consecuencia, un valor bajo del ratio significa, en principio, que la empresa no estará en condiciones de elegir, ya que los prestamistas entenderán que existe un alto riesgo de impago.

Un alto índice de endeudamiento puede motivar que la firma no tenga capacidad de adecuar las decisiones de inversión y financiación, no pudiendo elegir entre distintas acepciones financieras por haber perdido su autonomía con los riesgos asociados a esta pérdida.

Un esquema correcto de financiación debería permitir la cobertura del inmovilizado con capitales propios y con recursos exigibles a plazos por lo menos iguales al de su vida útil, para evitar problemas asociados a la refinanciación y además, atender a las necesidades del fondo de maniobra.

La autonomía financiera está ligada al apalancamiento financiero. Una entidad que tenga autonomía financiera puede optar a los beneficios

derivados del apalancamiento financiero positivo, hecho difícil de conseguir para una empresa que haya perdido su autonomía.

Además la independencia financiera también se haya unida a la solvencia, entendida ésta en sentido amplio; esto es, garantía suficiente por los activos y por la generación de flujos de tesorería para hacer frente a la remuneración de los capitales propios y ajenos y a la restitución de estos últimos.

La independencia financiera se alcanza cuando se da una relación óptima entre los recursos propios y ajenos. No obstante, la significación del ratio estará en función de la empresa a que se aplica y del sector donde opere; a nivel general, se suele situar en torno a 1, entre 0,7 y 1,5. Un valor superior a los citados transmite la imagen de que son los propietarios de la empresa los que verdaderamente soportan el riesgo que comporta su negocio. Este dato debe, no obstante, ser contrastado y aceptado o no por cada empresa en función de sus condiciones particulares.

El inverso de la tasa de autonomía financiera es *la tasa de endeudamiento* que informa sobre la relación existente entre el pasivo exigible total y el patrimonio neto.

*Ratio de exposición aseguradora*

$$R10 = \frac{\text{Provisiones Técnicas}}{\text{Capitales Propios}}$$

Es otro ratio de apalancamiento. Este ratio refleja el grado de apalancamiento de las obligaciones técnicas de la compañía en relación con los recursos propios. Mide el grado de absorción por la compañía de los errores de estimación en sus provisiones. También puede expresarse neto de reaseguro (Millán, 2000, p. 104).



Con este ratio se pretende medir la relación entre recursos administrados y recursos propios. Si bien su estudio se debe efectuar por comparación, un criterio de prudencia puede aconsejar que no sea muy elevado ya que significaría un volumen de recursos administrados (provisiones técnicas) muy superior al de los recursos propios, situándose la empresa en los límites de su capacidad de gestión, con menos margen de maniobra para enfrentarse a eventuales problemas de equilibrio financiero.

#### Ratio combinado y sus variantes

Los tratadistas del análisis contable aplicado al sector seguros han establecido, con carácter general, un indicador para evaluar la gestión global de la actividad aseguradora de nominado ratio combinado (Millán, 2000, p.146-148).

Se define de forma genérica como la suma de otros dos ratios:

- *Ratio de Siniestralidad* definido como

$$\frac{\text{Gastos Técnicos}}{\text{Ingresos adquiridos}}$$

- *Ratio Global de gastos de gestión* definido como:

$$\frac{\text{Gastos de Gestión}}{\text{Ingresos suscritos}}$$

Este ratio según se ha definido presenta dos grandes críticas:

- Se determina como suma de relaciones con distinto denominador: El mejor indicador de la calidad de la gestión de siniestros es el ratio de siniestralidad, ya que relaciona los ingresos de un período debidamente periodificados con sus gastos técnicos, por lo que es lógica su inclusión en el ratio combinado. Sin embargo, medir la

calidad de aplicación de los gastos de gestión es más complicado por su heterogeneidad. Habrá que buscar algún indicador adecuado que cumpla la condición de relacionar estos gastos con los ingresos que los ha generado, siempre teniendo en cuenta el reaseguro cedido. En consecuencia, en el numerador deben figurar los gastos de gestión del ejercicio netos de las comisiones de reaseguro y en el denominador, como ingresos que han generado dichos gastos, debe incluir los ingresos brutos no periodificados, puesto que su consumo principal (devengo de comisiones, gastos de agencia y apertura de pólizas), se efectúa en el momento de la suscripción y no a lo largo de la vida de la póliza.

- La segunda crítica viene derivada de que en su diseño no se recoge el reaseguro cedido. Para responder a esta cuestión se incorpora, como variante del ratio combinado el reaseguro cedido.

R15= Ratio de Siniestralidad de seguro directo +Ratio de Gastos

R16= Ratio de Siniestralidad de negocio neto +Ratio de Gastos Netos

### Ratios de liquidez

$$R17 = \frac{\text{Gastos técnicos directos} + \text{Comis. y otros gastos de explotación}}{\text{Primas y Recargos de seguro directo}}$$

$$R18 = \frac{\text{Gastos técnicos netos} + \text{Comis. y otros gastos de explotación}}{\text{Primas y Recargos de negocio neto}}$$

$$R20 = \frac{\text{Gastos técnicos directos}}{\text{Primas y Recargos de seguro directo}}$$

$$R21 = \frac{\text{Gastos técnicos netos}}{\text{Primas y Recargos de negocio neto}}$$

Se considera que pueden modificar la posición de solvencia a largo plazo. Aunque, en principio, informan de la capacidad de afrontar los gastos anuales con las primas anuales correspondientes, es decir, la vertiente de liquidez (Martín *et al.*, 1999, p. 78-79).

Un primer grupo de ratios de liquidez serían los ratios de gastos, llamados por la N.A.I.C., *Expense Ratios*, y son nuestros R17 y R18. En ellos se refleja qué proporción suponen los gastos de explotación respecto de las primas y recargos.

El **R17** es uno de los ratios utilizado por la NAIC (National Association Insurance Commissioners) en su sistema IRIS (Insurance Regulatory Information System) para detectar qué compañías aseguradoras pueden necesitar atención reguladora. Es el de nominado *Ratio combinado, base comercial* (Pozo, 1998, p. 37-48), que une los gastos por siniestros con los demás gastos en un único ratio al unir el ratio de siniestros con el ratio de gastos en base comercial.

Un segundo grupo de ratios de liquidez serían los ratios de siniestralidad, llamados *Loss Ratios* por la N.A.I.C. y se corresponden con el R20 y R21.

Con estos últimos dos ratios se obtiene la proporción que suponen los siniestros del año (riesgo incurrido) sobre las primas destinadas a la protección o cobertura de dichos riesgos. Esto es, en qué medida las primas financian los riesgos para los cuales estaban destinadas. Hay separación entre seguro directo y neto.

Estos dos últimos ratios, tanto el que considera el seguro directo o el neto, se denominan *ratio de siniestralidad* o *ratio de siniestralidad neta*. Ambos consideran en su numerador los gastos técnicos que son, en una compañía de seguros, los equivalentes al coste industrial de una empresa de

transformación, de ahí la importancia que presenta su análisis en el contexto de estas entidades (Millán, 2000, p.123). Es el indicador por excelencia de la gestión de los gastos técnicos; a través de su evolución y contrastación con el sector se puede determinar la calidad de la gestión técnica. Mediante su cotejo con la base técnica de la sociedad se puede apreciar el nivel de desviación respecto a las provisiones de coste y, en su caso, el déficit en los cálculos del precio de la póliza, permitiendo su reajuste a la mayor rapidez, siempre y cuando no esté regulado por la Administración. Esta regularización, en caso de existir deteriora a la larga el sentido y uso de este ratio.

En pocas palabras, la eficiencia económica de un sistema de seguros se mide por su ratio de siniestralidad que es el cociente entre primas y siniestros.

*Ratio de provisiones técnicas del reaseguro cedido sobre provisiones técnicas*

$$R19 = \frac{\text{Provisiones Técnicas de reaseguro cedido}}{\text{Provisiones Técnicas}}$$

Ya mencionamos en la introducción que en cualquier análisis de una empresa de seguros es necesario considerar una variable clave en la misma como es el reaseguro. El reaseguro es una práctica común en cualquier compañía aseguradora ya que es una forma de diversificar el riesgo, pero al mismo tiempo también es necesario que este no sea desproporcionado para evitar una excesiva dependencia del mismo, que en caso de insolvencia de la entidad reaseguradora, se comprometa la estabilidad de la aseguradora directa.

Es decir, las entidades aseguradoras tienen unas características técnico-financieras que le permiten cubrir las consecuencias de los riesgos

hasta un límite determinado (Linares, 1998, p.208). Cuando los riesgos contratados tienen una cuantía superior a este límite, la entidad sólo tiene capacidad para asumir una parte del mismo, estando técnicamente obligada a ceder el resto. Una forma de hacer frente a este problema es compartir la cuantía entre varios aseguradores; esta es la figura del *coaseguro*, pero otra forma es la de asumir frente al asegurado la totalidad del riesgo, asegurando a su vez con otra entidad, llamada *reasegurador*, la parte del riesgo no asumible, cubriendo sólo una parte de él, denominada *pleno de retención*. Cuando una entidad acepta ciertos riesgos y éstos exceden su pleno de retención, se ve obligada a su vez, a ceder los excesos a otras entidades. Esta figura se denomina *retrocesión*.

Una vez distinguida la figura del reaseguro de otras con las que guarda similitud, es necesario comprender, desde el punto de vista del análisis de la solvencia de una compañía de seguros, cómo funciona el reaseguro y cuál es su impacto en el balance y cuenta de resultados de la compañía. Describiremos brevemente en qué consiste.

El reaseguro es la operación por la cual una compañía de seguros transfiere a otra empresa aseguradora, denominada *reasegurador*, una parte de los riesgos del asegurado que ha aceptado y asumido, bien directamente por contratación directa (mediante suscripción de una póliza de seguro), o bien a través de otro asegurador por cesión de éste (*retrocesión*), a cambio de un precio o *prima de reaseguro* (Fernández, 2000).

Es un desplazamiento de una parte de los riesgos inicialmente transferidos por el asegurado a la compañía aseguradora (*asegurador directo*), quien a su vez los cede o transfiere en parte (de ahí que en estas operaciones se la denomine *empresa cedente* o también *reasegurado*) hacia otra empresa aseguradora (*reasegurador*), quien los acepta por la parte cedida (denominada *aceptante*).

De acuerdo con el Profesor Prieto (Prieto, 1973, p. 50) la necesidad del reaseguro para la entidad aseguradora puede tener diferentes orígenes, entre otros:

“ Cuando los siniestros acaecidos dan lugar a una indemnización total de gran cuantía, cosa que puede deberse a la no homogeneidad de la cartera desde el punto de vista de la suma asegurada o sencillamente al azar. La cuestión es que la cuantía de la indemnización de todos o parte de los siniestros acaecidos fue superior a la cuantía media de la distribución de la correspondiente variable aleatoria asociada a la cuantía de la indemnización por siniestro para la clase o clases de riesgos considerados... El ente asegurador busca, mediante el reaseguro, la cobertura de estos riesgos, en cuanto que no resulten absorbibles por el mismo y, en consecuencia, comprometan su supervivencia... El desequilibrio, en orden a los resultados ofrecidos por una cartera de seguros, puede aparecer por insuficiencia de las primas cobradas por la entidad aseguradora para la cobertura del riesgo. Aquí no consideramos este caso, admitiendo siempre primas equitativas, modeladas al riesgo cubierto.”

Por tanto, de igual manera que el asegurado transfiere riesgos al asegurador directo, éste realiza una transferencia o cesión de una parte de los riesgos que tiene asumidos por las pólizas de seguro suscritas al reasegurador, en función de su capacidad económica y financiera. Con ello cambia un coste de pérdida probable incierta, cuyo montante total para el conjunto de la cartera es de cuantía imprevisible, por un coste cierto y fijo determinado en la prima de reaseguro.

Cuando los riesgos son transferidos por un reasegurador a otro, se trata de una operación de *reaseguro retrocedido*, o simplemente *retrocesión* (al reasegurador se le llama retrocesionario).

La contratación de reaseguro se convierte de esta forma en uno de los objetivos principales de la compañía aseguradora para controlar y reducir las pérdidas económicas potenciales derivadas del negocio asegurador. Contribuye, por tanto, a reforzar la confianza de los asegurados en la capacidad de la compañía de hacer frente a grandes riesgos para los cuales ha demandado su cobertura.

Al hilo de esta última cuestión, y siguiendo al Profesor Prieto (Prieto, 1973, p. 53), el reaseguro supone:

“ Que la pérdidas o siniestros reasegurados sean en principio a cargo del asegurador directo. El asegurador directo se obliga a facilitar unas ciertas garantías a sus asegurados, de modo que, acaecido un siniestro, el asegurador hasta los límites de las garantías responde del pago de las correspondientes indemnizaciones. Si parte del riesgo asumido por el asegurador directo fuera cedido a uno o varios reaseguradores, esto no significaría otra cosa que el asegurador directo cubre una parte de las responsabilidades contraídas ante los asegurados, frente a los cuales continúa siendo el único responsable.... El reaseguro es una relación entre dos aseguradores, careciendo el reasegurador de relación inmediata con los asegurados, ante los cuales es responsable solamente el asegurador directo. Ante éste, lo son el reasegurador o reaseguradores, de acuerdo con las características del convenio o convenios del reaseguro establecidos.”

**ANEXO CAPITULO 4**



**TABLA 4.3. RATIOS UTILIZADOS EN EL ANÁLISIS**

RATIO	DEFINICIÓN	PARÁMETROS
R1	Fondo de maniobra/Activo Total $= (AC - PC) / \text{Activo Total}$ AC= Activo Circulante = Inversiones financ. + Créditos + Ajustes periodif.+ Efectivo + Prov. Reaseguro cedido PC= Prov. Prestaciones+ Deudas - (Empréstitos + Emp. Grupo + Emp.Asociadas) +Ajustes periodificación	Activo total = AA AC= A32+A4+A5+A6+A7 PC=P23+P5-(P51+P57+P58) +P6
R2	Beneficio antes de Impuestos(BAI)/Capitales propios (CP) CP = Capitales propios balance - Gastos Establecimiento	BAI= D6+D5 CP = P1- A21
R3	Ingresos Financieros/ Total Inversiones	$H3 / (A3 + A7)$
R4	BAI*/ Pasivo Total BAI* = BAI+ Amortizaciones + Provisiones + Rdos. Extraordinarios	$BAI^* = D6 + D5 + D2223 + D2224 + D35 + (D11121 - D11122) + (D11131 - D11132) + (H11122 - H11121) + (H11132 - H11131) + (H4 - D4)$
R5	Total Primas adquiridas de seguro directo / Capitales propios	$HD1 / (P1 - A21)$
R6	Total Primas adquiridas de negocio neto / Capitales propios	$H1 / (P1 - A21)$
R7	Total Primas adquiridas de seguro directo / Capitales propios + Prov. Técnicas	$HD1 / (P1 - A21 + P2)$
R8	Total Primas adquiridas de negocio neto / Capitales propios + Provisiones Técnicas	$H1 / (P1 - A21 + P2)$
R9	Capitales Propios / Pasivo Total	$(P1 - A21) / PP$
R10	Provisiones Técnicas / Capitales Propios	$P2 / (P1 - A21)$
R11	Gastos Técnicos de seguro directo / Capitales propios	$DD1 / (P1 - A21)$
R12	Gastos Técnicos de negocio neto / Capitales propios	$D1 / (P1 - A21)$

<b>R13</b>	Gastos Técnicos de seguro directo / Capitales propios + Prov. Técnicas	$DD1/(P1-A21+P2)$
<b>R14</b>	Gastos Técnicos de negocio neto / Capitales propios + Provisiones Técnicas	$D1/(P1-A21+P2)$
<b>R15</b>	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	RSD = $DD1/HD1$ RG = $D2/H2$
<b>R16</b>	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	RSN = $D1/H1$ RG = $D2/H2$
<b>R17</b>	(Gastos Técnicos de seguro directo + Comisiones y otros gastos de Explotación)/ Total Primas adquiridas de seguro directo	$(DD1+D2)/HD1$
<b>R18</b>	(Gastos Técnicos de negocio neto + Comisiones y otros gastos de Explotación)/ Total Primas adquiridas de negocio neto	$(D1+D2)/H1$
<b>R19</b>	Provisiones Técnicas de reaseguro cedido / Provisiones Técnicas	$A4/P2$
<b>R20</b>	RSD = Gastos Técnicos de seguro directo/ Total Primas adquiridas de seguro directo	RSD = $DD1/HD1$
<b>R21</b>	RSN = Gastos Técnicos de negocio neto/ Total Primas adquiridas de negocio neto	RSN = $D1/H1$

## **Capítulo 5**

# **EL MÉTODO ROUGH SET APLICADO A LA PREDICCIÓN DEL FRACASO EMPRESARIAL EN LAS EMPRESAS DE SEGUROS NO VIDA EN EL CASO ESPAÑOL**

## **I.- FUNDAMENTACIÓN DEL ANÁLISIS**

Se ha mencionado a lo largo de la exposición que en nuestra aplicación práctica vamos a seguir el enfoque original de la teoría Rough Set y no las extensiones de dicha teoría, que presuponen algún orden en el dominio de los atributos.

En nuestro trabajo vamos a utilizar ratios financieros. El nuevo enfoque permite tener en cuenta las propiedades de ordenación de los atributos analizados, si estas existen y son conocidas. Por ejemplo, si se sabe que un ratio financiero que toma un valor “alto” es preferible que tome un valor bajo, la firma con el valor del ratio alto se preferirá a la que tiene un valor del ratio bajo, y por tanto se considera que los valores del ratio se podrían ordenar (McKee, 2000, p.162).

La teoría convencional ha considerado a los ratios financieros como atributos que pueden ser ordenados. Por ejemplo, se considera que un ratio como el que relaciona el beneficio neto con el activo total es preferible que tome un valor alto a uno bajo.

Sin embargo cabe cuestionarse, si para los fines de predicción del fracaso empresarial, los ratios financieros puede considerarse que tengan un dominio ordenado. Un estudio de 1999 (McKee y Lensberg, 1999, p. 23) en el que se utiliza un algoritmo genético para desarrollar un modelo de predicción del fracaso empresarial encontró que la probabilidad de quiebra podría predecirse a través de un función de tres ratios financieros; activo total, beneficio neto/activo total, cash/pasivo circulante. Más aún, se encontró que el que un valor alto o bajo para un ratio determinado fuese considerado como bueno o malo dependía del valor que tomasen los otros dos ratios. En otras palabras, la teoría convencional de “cuanto mayor sea el valor de un ratio, mejor” quedó cuestionada ya que el estudio mostraba que una rentabilidad alta bajaba las probabilidades de quiebra, excepto cuando esa rentabilidad era inusualmente alta. En este último caso, un aumento en la rentabilidad aumentaba la probabilidad de

quiebra, excepto para el caso en que la empresa en cuestión fuese muy pequeña.

De acuerdo a este resultado, sería aconsejable, no asumir (al menos de forma generalizada), que los ratios financieros tienen una relación de dominio. De acuerdo con esto, nuestra investigación utiliza el enfoque original de la teoría Rough Set basado en la relación de no diferenciación.

## **II. APLICACIÓN DEL ANÁLISIS Y RESULTADOS**

### **II.1.- Definición Del Problema**

En otras secciones hemos explicado como la Teoría Rough Set puede analizar muy diversos problemas de decisión con atributos múltiples. De hecho, está especialmente indicada para un tipo de problemas multiatributo que son los problemas de clasificación.

Dentro de los problemas de clasificación hemos comentado la existencia de problemas de clasificación multiatributo consistentes en la asignación de cada objeto (descrito por un conjunto de valores de atributos) a una clase o categoría previamente establecida. La predicción del fracaso empresarial es un ejemplo claro de este tipo de problemas ya que pretendemos asignar empresas (nuestros objetos), descritas por los valores de un conjunto de ratios financieros (nuestros atributos), a una clase o categoría (sana o fracasada).

Probaremos la Teoría Rough Set, descrita en las secciones previas sobre un conjunto de datos reales extraídos de entidades aseguradoras españolas del ramo no-vida. Nuestro objetivo es en primer lugar reducir el conjunto de ratios financieros introducidos en el modelo eliminando aquellos que no son esenciales (innecesarios) para caracterizar una empresa como sana o fracasada. Buscamos el mínimo conjunto de atributos para poder establecer un “sistema de alerta temprana de insolvencia” para entidades aseguradoras que redujese el tiempo y los costes empleados por los usuarios para evaluar el problema de la insolvencia en una entidad aseguradora.

## II.2.- Aplicación De La Metodología Rough Set

### Tabla de Información

Para la aplicación de esta metodología lo primero que hemos construido es una tabla de información con las 72 empresas, 36 sanas y 36 fracasadas. Las empresas han sido evaluadas de acuerdo con los valores que toman los 21 ratios elegidos para cada una de ellas. Estas empresas han sido clasificadas a una de las dos categorías, sana o fracasada, para el año que denominaremos *base* (año 1), que es el año anterior a la quiebra. Por tanto estamos considerando un atributo de decisión,  $d$ , que efectúa en esta tabla de información (a partir de ahora tabla de decisión) una partición dicotómica: si  $d=0$  la firma pertenece al grupo de las fracasadas y si  $d=1$ , la empresa pertenece al grupo de las sanas.

Hemos de mencionar que aunque se calcularon los ratios 15 y 16, no han sido utilizados en el análisis posterior porque las peculiaridades de la muestra hacían que tomaran valores que carecían de sentido económico, al no tener la mayoría de las empresas otros ingresos de explotación. Con lo que solo hemos utilizado los 19 restantes en nuestro sistema de información.

Hemos efectuado previamente una *análisis descriptivo* de los ratios obteniendo de todos ellos las principales medidas estadísticas (media, varianza desviación típica, rango, valores máximos, valores mínimos, coeficiente de asimetría y de curtosis, percentiles; ver ANEXO CAPÍTULO 5).

No hemos calculado la matriz de correlación para el conjunto de ratios, aunque sabemos que algunos de ellos están muy correlacionados por dos motivos:

1.- No es un modelo estadístico con lo que los resultados no se ven afectados por las posibles correlaciones entre las variables. Es más, al poder incluir todas las variables no perdemos el contenido informativo que puede existir en ratios semiindependientes aumentando así el poder predictivo, principal objetivo de la construcción del modelo.

2.- Dado que la metodología rough set proporciona el/los mínimo/s conjunto/s de variables que dan la misma capacidad de clasificación que el conjunto de todas ellas, es decir, el/los reducto/s, en un mismo reducto no debería haber ratios muy correlacionados y si los hubiera, al hacer un análisis de los mismos descartaríamos aquellos que contuvieran ratios muy correlacionados.

En consecuencia, una vez seleccionados de entre todos los reductos, algunos que a nuestro juicio contienen los atributos más significativos, para poder elegir el reducto definitivo se han calculado algunos coeficientes de correlación y así asegurarnos de que el reducto que elijamos no contenga ratios muy correlacionados. Por tanto, en algunos casos si hemos calculado el coeficiente de correlación aunque no la matriz completa.

Después, hemos calculado las *medias* de cada ratio en los cinco años considerados para ver su evolución en este período, distinguiendo entre empresas sanas y fracasadas, al objeto de comprobar el comportamiento de los ratios y sus posibles diferencias entre los casos de fracaso y no fracaso. Es decir, el análisis a través de las medias de cada variable en un período de tiempo, puede dar una idea de la evolución y diferenciación, entre los ratios de las empresas consideradas sanas, respecto a los de aquellas que fracasaron por el motivo que fuese.

En la tabla 5.1 se indican los valores de las medias:



**TABLA 5.1.- MEDIAS PARA LOS DISTINTOS HORIZONTES**

Ratios	Horizonte	1 Año	2 Años	3 Años	4 Años	5 Años
	Tipo Empre.					
R1	Sanas	0.39	0.62	0.41	0.38	0.37
	Fracasadas	0.15	0.23	0.24	0.26	0.23
R2	Sanas	0.06	0.06	0.09	0.09	0.10
	Fracasadas	0.08	-0.23	0.02	0.06	0.07
R3	Sanas	0.1	0.08	0.09	0.10	0.10
	Fracasadas	0.09	0.07	0.12	0.18	0.07
R4	Sanas	0.11	0.12	0.13	0.14	0.13
	Fracasadas	0.20	0.18	0.24	0.14	0.20
R5	Sanas	2.12	-3.42	2.64	2.49	3.66
	Fracasadas	38.7	8.59	2.61	2.39	10.05
R6	Sanas	1.69	-1.65	2.20	2.16	3.23
	Fracasadas	37.08	5.64	2.33	2.11	1.96
R7	Sanas	0.79	0.88	1.06	1.02	1.10
	Fracasadas	2.94	1.25	0.88	0.85	8.94
R8	Sanas	0.78	0.87	1.04	1.01	1.09
	Fracasadas	2.53	1.25	0.88	0.88	1.11
R9	Sanas	0.49	0.49	0.43	0.42	0.38
	Fracasadas	0.29	0.39	0.37	0.39	0.41
R10	Sanas	2.05	-4.33	2.12	2.03	2.67
	Fracasadas	45.98	8.04	2.27	1.92	1.19
R11	Sanas	1.49	-4.00	1.66	1.58	2.28
	Fracasadas	34.74	7.23	1.98	1.77	2.55
R12	Sanas	1.12	-2.11	1.29	1.32	1.94
	Fracasadas	34.26	5.10	1.80	1.44	1.32
R13	Sanas	0.49	0.54	0.59	0.59	0.63
	Fracasadas	1.95	0.90	0.67	0.61	1.93
R14	Sanas	0.47	0.52	0.56	0.58	0.61
	Fracasadas	1.59	0.89	0.68	0.59	0.75
R17	Sanas	1.12	1.30	1.28	1.28	1.05
	Fracasadas	1.39	1.18	1.19	1.11	1.43
R18	Sanas	1.18	1.35	1.34	1.42	1.10
	Fracasadas	1.42	1.22	1.25	1.26	2.52
R19	Sanas	0.16	0.15	0.17	0.17	0.16
	Fracasadas	0.12	0.12	0.15	0.20	0.18
R20	Sanas	0.63	0.74	0.86	0.62	0.62
	Fracasadas	0.66	0.72	0.74	0.67	0.63
R21	Sanas	0.62	0.72	0.83	0.61	0.62
	Fracasadas	0.62	0.70	0.73	0.68	0.65

	<b>Nº Empresas Sanas</b>	<b>Nº Empresas Fracasadas</b>
<b>Año 1</b>	36	36
<b>Año 2</b>	35	35
<b>Año 3</b>	27	30
<b>Año 4</b>	24	27
<b>Año5</b>	19	21

El utilizar la media como medida representativa de una muestra puede causar problemas de significación, cuando la dispersión en torno a dicha media es grande, hecho que puede producirse al aparecer algunos valores atípicos en la muestra. En nuestra muestra en el cálculo de los valores de los ratios aparecieron algunos valores extremos o atípicos, que son conocidos con la denominación de valores *outliers*. Dichos valores no han sido eliminados en la aplicación del método rough set porque su valor se ha discretizado al igual que para el resto de los datos y porque una de las ventajas de este enfoque es que toma la información tal cual se presenta en la realidad, con sus inconsistencias y sin que sea necesario que los datos verifiquen ninguna hipótesis como las que se requieren para aplicar la mayoría de las técnicas estadísticas. Sin embargo la presencia de estos valores atípicos, sí afecta al cálculo de algunas de las medias, que están distorsionadas por la presencia de los mismos.

Del análisis de la evolución de las medias, y pese al inconveniente de la presencia de outliers, se deduce que para la mayoría de los ratios, las medias presentan mejores valores para las empresas sanas que para las fracasadas, y en éstas los valores se van deteriorando a medida que se acerca el fracaso. Esto demuestra que la capacidad de los ratios para distinguir entre empresas sanas y fracasadas va disminuyendo a medida que nos alejamos del momento de la crisis. Así para el año 1 antes del

fracaso los ratios presentan en la mayoría de los casos valores muy diferenciados entre empresas sanas y fracasadas, diferencia que va disminuyendo para los años posteriores.

Por consiguiente, y como es lógico de esperar, una conclusión que se puede extraer del análisis del nivel de significación de los ratios es que a mayor cercanía al momento del fracaso, el poder diferenciador de los ratios es superior.

En general, y tal como hemos indicado, de los diecinueve ratios considerados, se aprecian diferencias entre unas y otras empresas, a favor de las sanas. Los ratios más significativos son el R1, R2, R4, R9, R11, R12, R13 y R14 y donde más se decanta es en el R1 que es un ratio de liquidez. En consecuencia del análisis de la significatividad de los ratios en la muestra podemos extraer estas conclusiones de cara al análisis de la solvencia:

- *La importancia de la liquidez (R1)*: Se aprecia una mayor liquidez en las empresas sanas que en las que se van a liquidar en todos y cada uno de los años, lo cual indica problemas de liquidez desde mucho antes de la liquidación.

- *La necesidad de obtener una adecuada rentabilidad financiera (R3)* no solo por razones de imagen que le permitan colocar las ampliaciones de capital o conseguir créditos, sino también porque un beneficio suficiente en una firma con autonomía financiera lleva consigo el mantenimiento de la misma.

- *Una conveniente política de financiación (R4)*. Una financiación escasa debida a ingresos insuficientes que aparezcan de manera inesperada puede ser la causa del comienzo de proceso de quiebra.

- *Autonomía Financiera (R9)*: Las empresas sanas muestran una mayor autonomía financiera lo cual les posibilita el beneficiarse del

apalancamiento financiero y muestran por ello un mayor grado de este último.

- *La importancia de evaluar la solvencia estática y dinámica (R11, R12, R13, R14).*

Además algunos de ellos también son indicativos de:

- Una mala fijación de las primas: La inversión del proceso productivo del sector asegurador, hace que el proceso de tarificación sea clave. Este proceso no es una ciencia exacta y primas adecuadamente estimadas en el pasado pueden no estarlo en el presente si las condiciones cambian y llevan consigo un aumento de los siniestros. Es evidente que si la mala tarificación da como resultado una insuficiencia de la prima para cubrir los siniestros, este hecho llevaría a resultados técnicos negativos de difícil absorción por los resultados financieros, e incluso por los capitales libres a la larga. Este factor gana en importancia según nos alejamos al momento de la crisis.

- Una infradotación de las provisiones que lleva consigo a que el beneficio del asegurador no sea el real si no que esté distorsionado y en concreto sea superior al real. Esto conduce a que dicho beneficio pueda ser repartido con la consecuente descapitalización de la empresa o a que se pueda aumentar más el volumen de negocio de lo que en realidad se debería si el beneficio no estuviese sobreestimado y con el lógico incremento en los riesgos que no estaría cubierto.

- Las empresas fracasadas concentran un mayor volumen de riesgo con vencimiento en el ejercicio siguiente al de cobro de las primas.

### Discretización

Antes de comenzar la aplicación hemos de adecuar la información que contiene nuestro sistema de información.

En las aplicaciones prácticas, los sistemas de información a menudo contienen datos de carácter continuo y discreto. El análisis rough set de los sistemas de información da buenos resultados cuando los dominios de los atributos son conjuntos finitos cuyo número cardinal es relativamente bajo. Este requisito a menudo se satisface cuando los atributos tienen un carácter cualitativo (discreto). Sin embargo, los atributos que toman valores arbitrarios de unos intervalos dados, es decir, tienen un carácter cuantitativo, deben ser tratados para traducir sus valores a términos cualitativos (por ejemplo bajo, medio y alto) utilizando los correspondientes valores numéricos como "1, 2, 3". Esta manipulación implica una división del dominio original en algunos subintervalos y la asignación de códigos cualitativos a dichos subintervalos.

No existe un modo general de definir de manera óptima los valores frontera de los subintervalos. La definición de los valores frontera de los subintervalos debería tener en cuenta la experiencia, el conocimiento, los hábitos y costumbres utilizados por los expertos e, incluso, un error de medida en algunos atributos. Muchas veces los valores frontera surgen normalmente de convencionalismos, hábitos o de asignación subjetiva.

Es necesario resaltar que la definición de los valores frontera de los subintervalos puede influir en los resultados, y en particular, en la calidad de la clasificación, por tanto se hace necesario revisar la sensibilidad de los resultados frente a cambios en la definición de estos valores.

Por tanto, vamos a traducir los valores cuantitativos de los ratios financieros a términos cualitativos. Por esta razón vamos a considerar dos tipos de sistema de información: *el sistema de información original* y *el sistema de información codificado*. El primero contiene los ratios

financieros (atributos cuantitativos) con los valores de sus dominios originales. El segundo contiene atributos con valores codificados. La teoría rough set analiza posteriormente el sistema de información codificada.

La codificación de los valores de los ratios la hemos efectuado sustituyendo los valores originales por códigos que, en nuestro caso, son números que designan a los subintervalos que los contienen. Es decir, se han codificado utilizando la denominada *tabla de normas*.

La tabla de normas es una estructura que contiene las definiciones de todos los valores frontera de los subintervalos definidos por el usuario.

En consecuencia, para cada ratio se ha buscado en cada grupo de empresas (sanas y fracasadas) cuales son sus niveles más representativos, con el objetivo final de establecer varios niveles de corte en los valores del ratio, que separen distintos estados de equilibrio o desequilibrio respecto a ese ratio.

Para el establecimiento de esos puntos de separación, es decir, las *normas*, hemos calculado los cuartiles de cada ratio. Es decir, los percentiles 25, 50, y 75 de cada ratio nos informan, respectivamente, de las empresas que toman un valor para dicha variable por debajo de ese percentil. Además la utilización de los cuartiles evita cualquier subjetividad y no excluye la posibilidad de que la experiencia o conocimiento de un experto modifiquen dichos valores. La tabla 5.2. muestra los valores utilizados para discretizar los distintos ratios.

La utilización de percentiles para establecer intervalos en los ratios es frecuente en los trabajos de investigación. Podemos citar otros investigaciones que utilizan este procedimiento: Laitinen (1992), García *et al.*, (1997) o McKee, (2000).

**TABLA 5.2. INTERVALOS**

Ratio	1 <sup>st</sup>	2 <sup>nd</sup>	3 <sup>rd</sup>	4 <sup>th</sup>
R1	$(-\infty, 0.115]$	$(0.115, 0.295]$	$(0.295, 0.475]$	$(0.475, +\infty)$
R2	$(-\infty, 0]$	$(0, 0.1]$	$(0.1, 0.07]$	$(0.07, +\infty)$
R3	$(-\infty, 0.03]$	$(0.03, 0.06]$	$(0.06, 0.11]$	$(0.11, +\infty)$
R4	$(-\infty, 0.03]$	$(0.03, 0.08]$	$(0.08, 0.26]$	$(0.26, +\infty)$
R5	$(-\infty, 0.565]$	$(0.565, 1.565]$	$(1.565, 3.29]$	$(3.29, +\infty)$
R6	$(-\infty, 0.525]$	$(0.525, 1.38]$	$(1.38, 2.715]$	$(2.715, +\infty)$
R7	$(-\infty, 0.455]$	$(0.455, 0.725]$	$(0.725, 1.22]$	$(1.22, +\infty)$
R8	$(-\infty, 0.46]$	$(0.46, 0.7]$	$(0.7, 1.18]$	$(1.18, +\infty)$
R9	$(-\infty, 0.14]$	$(0.14, 0.35]$	$(0.35, 0.68]$	$(0.68, +\infty)$
R10	$(-\infty, 0.04]$	$(0.04, 0.545]$	$(0.545, 2.97]$	$(2.97, +\infty)$
R11	$(-\infty, 0.27]$	$(0.27, 1.095]$	$(1.095, 2.43]$	$(2.43, +\infty)$
R12	$(-\infty, 0.27]$	$(0.27, 0.845]$	$(0.845, 1.815]$	$(1.815, +\infty)$
R13	$(-\infty, 0.27]$	$(0.27, 0.49]$	$(0.49, 0.82]$	$(0.82, +\infty)$
R14	$(-\infty, 0.225]$	$(0.225, 0.435]$	$(0.435, 0.765]$	$(0.765, +\infty)$
R17	$(-\infty, 0.98]$	$(0.98, 1.055]$	$(1.055, 1.27]$	$(1.27, +\infty)$
R18	$(-\infty, 1]$	$(1, 1.09]$	$(1.09, 1.29]$	$(1.29, +\infty)$
R19	$(-\infty, 0]$	$(0, 0.065]$	$(0.065, 0.19]$	$(0.19, +\infty)$
R20	$(-\infty, 0.515]$	$(0.515, 0.68]$	$(0.68, 0.785]$	$(0.785, +\infty)$
R21	$(-\infty, 0.515]$	$(0.515, 0.655]$	$(0.655, 0.75]$	$(0.75, +\infty)$

Hemos de resaltar que la utilización de *normas* que transformen los atributos cuantitativos en términos cualitativos no se impone por el rough set pero sí para una interpretación práctica de los atributos cuantitativos. Incluso si un atributo representa una medida continua, como es un ratio financiero, o la presión sanguínea en medicina, el experto normalmente

interpreta los valores de los atributos en términos cualitativos, esto es, bajo, medio o alto. Es más, este proceso se hace necesario debido a que la precisión de los ratios financieros es bastante dudosa, y dificulta el extraer conclusiones generales de los datos en términos de dependencias, reductos o reglas de decisión.

Además, y dado que las normas se utilizan desde el principio del problema y hasta el final en la explicación de las reglas de decisión, no falsean la imagen original de la situación.

Las empresas para las que los valores de estos ratios financieros están en los mismos subintervalos se considera que tienen el mismo comportamiento y características. Además, los códigos que se utilizan para referirse a un subintervalo no representan ningún orden de preferencias. La teoría rough set no tiene en cuenta el orden, ya que trabaja con la relación de no diferenciación, es más la selección de los códigos no afecta a los resultados obtenidos.

Una vez determinados los intervalos hay que asignar los códigos a los mismos.

Con la tabla anterior y el principio “cuánto más alto sea el código, mejor subintervalo”, tenemos las bases para obtener la tabla de asignación de códigos. Dado el principio que hemos aplicado, para los atributos decrecientes (aquellos en que los valores más bajos son mejores) se les ha asignado un código en orden inverso. Además, hemos efectuado algunas correcciones en la escala en el caso en que, nuestro juicio y/o preferencias, no estuvieran en concordancia con la secuencia de aumento o disminución de los subintervalos.

La siguiente tabla muestra la asignación de códigos a los subintervalos:



**TABLA 5.3. CÓDIGOS ASIGNADOS A LOS INTERVALOS**

Ratios	Intervalos			
	1º	2º	3º	4º
R1	1	2	3	4
R2	1	2	3	4
R3	1	2	3	4
R4	1	2	3	4
R5	1	3	4	2
R6	1	3	4	2
R7	1	3	4	2
R8	1	3	4	2
R9	1	3	4	2
R10	1	3	4	2
R11	1	4	3	2
R12	1	4	3	2
R13	1	4	3	2
R14	1	4	3	2
R17	1	4	3	2
R18	1	4	3	2
R19	1	3	3	2
R20	4	3	2	1
R21	4	3	2	1

ROSE software permite que el usuario introduzca la tabla de normas para crear automáticamente un fichero de entrada con la información codificada. Por tanto, los intervalos se han introducido en el ROSE que transformó automáticamente los valores de las empresas para los 19 ratios, obteniendo así la tabla de información codificada.

**TABLA 5.4. EMPRESAS FRACASADAS**

FIRMS	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R17	R18	R19	R20	R21	D
E1	3	1	3	3	4	4	4	4	3	4	3	3	4	3	4	3	3	3	2	0
E2	2	1	2	2	3	3	4	2	4	3	4	3	2	2	4	1	1	2	2	0
E3	1	1	1	2	3	3	4	4	3	3	4	3	2	2	3	3	1	1	1	0
E4	2	3	1	3	2	2	4	4	3	2	2	2	3	3	4	4	3	3	3	0
E5	1	2	3	1	1	1	1	1	2	3	1	1	1	1	2	2	2	4	4	0
E6	1	1	3	4	1	1	4	4	1	1	1	1	3	3	2	2	3	2	1	0
E7	1	4	1	4	2	2	2	2	1	2	2	2	3	2	1	1	3	3	3	0
E8	2	1	3	4	2	2	4	4	1	2	2	2	3	3	3	4	3	2	2	0
E9	3	1	2	4	2	2	4	4	1	2	2	2	3	3	4	4	3	2	2	0
E10	1	1	1	4	1	1	4	4	1	1	1	1	3	3	3	3	1	2	2	0
E11	1	1	1	2	1	1	4	4	1	1	1	1	3	3	4	4	3	3	3	0
E12	1	1	4	4	1	1	2	2	1	1	1	1	2	2	3	3	1	3	3	0
E13	2	1	2	2	2	2	4	4	1	2	2	2	2	3	2	3	2	1	3	0
E14	2	4	1	3	2	2	4	4	3	2	3	3	4	4	1	1	3	3	3	0
E15	2	3	4	3	2	4	4	3	1	2	2	3	3	3	4	3	2	3	2	0
E16	1	1	4	1	1	1	2	1	1	1	1	1	2	1	2	2	3	1	1	0
E17	1	1	2	4	1	1	2	2	1	1	1	1	2	2	2	2	3	1	1	0
E18	2	4	4	3	1	1	1	1	4	3	1	1	1	1	2	2	3	2	2	0
E19	1	1	1	4	2	2	4	4	1	2	2	2	3	3	2	2	3	1	1	0
E20	2	1	2	3	4	4	1	1	3	4	3	3	4	4	3	3	3	1	1	0
E21	3	1	4	4	3	3	3	4	4	3	1	1	1	1	2	2	1	4	4	0
E22	1	2	1	1	1	1	1	1	4	1	1	1	1	1	1	1	1	4	4	0
E23	3	4	2	4	4	2	2	2	4	3	3	2	2	2	1	1	1	2	2	0
E24	2	1	2	2	3	3	3	3	3	4	4	4	4	4	1	4	2	4	3	0
E25	2	3	4	1	4	4	1	3	1	2	3	4	4	1	2	3	2	1	4	0
E26	4	4	3	3	4	4	2	2	2	3	4	3	2	2	1	1	1	3	2	0
E27	2	1	2	2	2	2	2	2	4	3	2	2	2	2	4	4	1	2	2	0
E28	1	2	1	1	1	3	3	3	4	1	4	4	1	4	1	1	1	3	3	0
E29	4	1	4	1	1	1	1	1	2	3	1	1	1	4	3	3	2	2	1	0
E30	4	4	1	3	4	2	2	2	4	4	3	2	2	2	1	1	1	2	2	0
E31	2	1	1	4	1	1	1	1	2	1	1	1	1	1	2	2	1	4	4	0
E32	2	1	2	3	3	3	3	1	4	4	4	4	4	4	2	2	2	2	3	0
E33	4	1	1	1	2	2	2	2	4	1	2	2	2	2	4	4	1	3	3	0
E34	1	3	1	3	2	2	2	2	4	4	2	2	2	2	4	1	1	4	4	0
E35	3	4	4	3	4	4	2	2	2	3	3	3	2	2	1	1	1	3	3	0
E36	4	3	2	2	4	4	2	2	2	3	3	3	2	2	1	1	1	1	1	0

**TABLA 5.5. EMPRESAS SANAS**

FIRMS	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13	R14	R17	R18	R19	R20	R21	D
E101	4	3	3	1	1	1	1	1	3	4	4	4	1	1	2	2	3	1	1	1
E102	4	1	3	1	3	3	3	3	2	1	1	1	1	1	3	4	1	4	4	1
E103	4	4	3	4	3	3	4	4	2	1	4	4	3	3	1	1	1	4	4	1
E104	3	4	2	3	4	4	1	3	1	2	3	3	1	4	3	4	2	2	3	1
E105	4	4	3	4	3	3	3	3	4	4	4	4	1	1	1	1	2	4	4	1
E106	3	1	4	2	2	2	4	4	1	2	2	2	3	3	4	3	2	1	1	1
E107	4	4	3	3	4	4	3	3	3	4	3	3	4	4	1	4	3	3	2	1
E108	3	1	4	4	2	2	3	3	1	2	2	2	3	3	2	2	3	1	1	1
E109	4	3	3	1	4	3	3	1	3	2	3	4	4	1	1	3	2	3	3	1
E110	3	3	3	1	3	3	1	1	3	4	4	4	4	4	2	3	2	1	2	1
E111	3	3	4	3	2	2	3	3	1	2	2	2	4	3	3	3	3	1	1	1
E112	3	4	4	3	4	4	2	2	4	3	3	2	2	2	3	3	1	1	1	1
E113	4	3	4	1	3	3	1	1	3	4	3	4	4	4	3	2	2	1	1	1
E114	2	1	3	2	4	4	3	3	3	2	2	3	3	4	3	3	3	1	2	1
E115	1	1	2	1	4	4	3	3	3	4	3	3	4	3	2	2	3	1	1	1
E116	3	1	2	4	2	2	4	4	1	2	2	2	3	3	3	3	3	2	2	1
E117	3	1	3	4	4	4	1	1	3	2	3	3	4	4	2	2	3	1	1	1
E118	4	4	3	3	4	4	2	2	4	4	4	4	4	4	1	1	1	4	4	1
E119	3	1	4	2	2	4	3	3	1	2	2	3	3	4	1	2	2	2	3	1
E120	3	1	3	1	3	3	3	3	3	4	4	4	4	4	3	3	3	2	2	1
E121	2	1	1	2	2	2	2	2	4	4	2	2	2	2	4	4	1	4	4	1
E122	2	2	3	1	1	1	3	3	2	1	1	1	1	1	3	4	1	4	4	1
E123	4	3	2	2	3	3	4	4	2	3	4	4	3	3	4	4	1	3	3	1
E124	2	2	3	1	4	4	1	3	3	4	3	3	4	4	3	3	2	3	3	
E125	1	3	2	2	3	3	3	3	3	4	4	4	1	1	1	1	3	4	4	1
E126	4	3	2	2	3	3	3	3	2	3	4	4	4	4	4	4	1	2	2	1
E127	3	4	3	4	4	4	2	2	4	3	3	3	3	2	4	4	1	3	3	1
E128	1	1	1	3	3	3	4	4	2	1	4	4	3	3	3	3	1	2	2	1
E129	3	2	1	2	3	3	3	3	4	3	4	4	1	4	1	4	3	4	4	1
E130	1	2	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1	4	4	1
E131	4	3	2	1	3	4	2	2	2	1	3	3	2	2	4	4	1	2	1	1
E132	3	3	3	2	3	1	1	1	4	4	4	4	1	1	3	2	2	2	2	1
E133	3	3	1	1	2	2	2	2	2	3	3	2	2	2	4	4	1	3	3	1
E134	1	2	1	1	1	1	1	1	2	1	1	1	1	1	4	1	1	4	4	1
E135	2	1	1	2	1	1	1	1	2	1	1	1	1	1	2	2	3	4	4	1
E136	4	1	1	1	1	1	1	1	2	1	1	1	1	1	2	2	1	4	4	1

La tabla codificada, que es la que se hemos utilizado para un posterior análisis, constaba de las 72 empresas descritas por los 19 atributos codificados (ratios financieros), utilizando los datos del año previo a la quiebra (año -1) y asignadas de forma binaria a una clase de decisión (sana o fracasada, representado por 1 ó 0 respectivamente).

Ya se ha indicado que el software ROSE calculó de manera automática la tabla codificada, y es necesario comentar que todo el análisis rough set de dicha tabla se ha hecho siguiendo dicho software.

### Experimentos Realizados

Hemos realizado dos tipos de experimentos, obteniendo de cada uno de ellos dos modelos, por lo que nos referiremos a ellos como **modelo 1** y **modelo 2**.

Las reglas de decisión del modelo 1 se han obtenido empleando los datos del año 1 (año anterior a la quiebra) y verificando la capacidad predictiva de las mismas utilizando los datos de los años 2, 3, 4 y 5 previos a la quiebra. Posteriormente se ha efectuado un análisis discriminante siguiendo los mismos criterios que para el análisis rough set, es decir, estimar la función discriminante con el año 1 y comprobar la capacidad predictiva de dicha función utilizando los años 2, 3, 4 y 5 anteriores a que las firmas quebraran.

Dado que utilizar los mismas empresas para validar y para estimar puede condicionar los resultados obtenidos hemos obtenido un segundo modelo siguiendo este procedimiento: De las 72 empresas de que consta la muestra original (36 sanas y 36 fracasadas), y para el año 1 (año previo a la quiebra) hemos utilizado únicamente 27 empresas de cada una de las submuestras para estimar las reglas de decisión, dejando las 9 restantes para comprobar la validez de las mismas. Pretendemos de esta forma aplicar el modelo a datos que no hubieran sido utilizados para la elaboración del mismo. En consecuencia, en este segundo modelo tendremos una muestra de *entrenamiento* para obtener las reglas formada por 54 empresas y una muestra de *validación* para verificar el modelo formada por 18 empresas. La selección de la muestra de validación se ha efectuado aleatoriamente tomando las empresas numeradas en Sanchis (2000, 2002) de la 19 a la 27, para las fracasadas, y de la 119 a la 127 para las sanas como submuestras de validación.

Dado que muchos resultados son comunes a ambos experimentos los iremos comentando conjuntamente y refiriéndonos a ellos como modelo 1 y modelo 2 en el caso en que sean comentarios específicos a cada uno de ellos.

### Precisión y Calidad de la Aproximación

El primer resultado que obtenemos con el análisis rough set es la aproximación para cada una de las clases que estamos considerando, sana y fracasada, por el conjunto de atributos (ratios financieros). En ambos modelos, y como puede comprobarse en las tablas 5.6. Y 5.7., la totalidad del conjunto de atributos dio una aproximación perfecta de las clases de decisión, esto es igual a uno. Además y, en consecuencia, la calidad de la clasificación es también igual a uno.

**TABLA 5.6. MODELO 1**

<b>Clases</b>	<b>Número de Empresas</b>	<b>Aproximación por debajo</b>	<b>Aproximación por arriba</b>	<b>Precisión</b>
0 (fracasadas)	36	36	36	1
1 (sanas)	36	36	36	1

**TABLA 5.7. MODELO 2**

<b>Clases</b>	<b>Número de Empresas</b>	<b>Aproximación por debajo</b>	<b>Aproximación por arriba</b>	<b>Precisión</b>
0 (fracasadas)	27	27	27	1
1 (sanas)	27	27	27	1

El resultado es muy significativo porque a pesar de tener un gran sistema de información, las empresas están muy bien discriminadas entre ellas.

### Núcleo y Reductos

El siguiente paso del análisis es la construcción del mínimo subconjunto de atributos independientes que aseguren la misma calidad de clasificación que la totalidad del conjunto de todos ellos, esto es los reductos, a partir del sistema de información codificado.

Para el primer modelo hemos obtenido 241 reductos que son:

- |                                 |                                |
|---------------------------------|--------------------------------|
| 1: {R2,R3,R4,R6,R17,R18,R21}    | 39: {R4,R8,R9,R17,R18,R19,R21} |
| 2: {R2,R3,R4,R5,R17,R18,R21}    | 40: {R4,R8,R9,R14,R18,R19,R21} |
| 3: {R2,R3,R4,R6,R17,R19,R21}    | 41: {R4,R8,R9,R13,R18,R19,R21} |
| 4: {R2,R3,R4,R5,R17,R19,R21}    | 42: {R4,R8,R9,R12,R18,R19,R21} |
| 5: {R1,R2,R3,R4,R17,R19,R21}    | 43: {R4,R8,R9,R11,R18,R19,R21} |
| 6: {R1,R3,R9,R17,R19,R20}       | 44: {R4,R6,R8,R9,R18,R19,R21}  |
| 7: {R1,R3,R9,R14,R17,R19}       | 45: {R4,R5,R8,R9,R18,R19,R21}  |
| 8: {R1,R3,R9,R13,R17,R19}       | 46: {R2,R4,R8,R9,R18,R19,R21}  |
| 9: {R1,R3,R9,R10,R17,R19}       | 47: {R4,R7,R9,R17,R18,R19,R21} |
| 10: {R1,R3,R8,R9,R17,R19}       | 48: {R4,R7,R9,R14,R18,R19,R21} |
| 11: {R1,R3,R7,R9,R17,R19}       | 49: {R4,R7,R9,R13,R18,R19,R21} |
| 12: {R1,R3,R6,R9,R17,R19}       | 50: {R4,R7,R9,R12,R18,R19,R21} |
| 13: {R1,R3,R5,R9,R17,R19}       | 51: {R4,R7,R9,R11,R18,R19,R21} |
| 14: {R1,R3,R9,R11,R13,R18,R19}  | 52: {R4,R6,R7,R9,R18,R19,R21}  |
| 15: {R1,R3,R11,R17,R18,R19,R21} | 53: {R4,R5,R7,R9,R18,R19,R21}  |
| 16: {R1,R3,R8,R11,R17,R19,R21}  | 54: {R2,R4,R7,R9,R18,R19,R21}  |
| 17: {R1,R3,R7,R11,R17,R19,R21}  | 55: {R3,R4,R11,R17,R18,R21}    |
| 18: {R1,R3,R11,R13,R17,R18,R19} | 56: {R3,R4,R11,R12,R17,R21}    |
| 19: {R1,R3,R8,R11,R13,R17,R19}  | 57: {R3,R4,R8,R11,R17,R21}     |
| 20: {R1,R3,R7,R11,R13,R17,R19}  | 58: {R3,R4,R7,R11,R17,R21}     |
| 21: {R1,R3,R10,R11,R17,R18,R19} | 59: {R2,R3,R4,R11,R17,R21}     |
| 22: {R1,R3,R8,R10,R11,R17,R19}  | 60: {R1,R3,R4,R11,R17,R21}     |
| 23: {R1,R3,R7,R10,R11,R17,R19}  | 61: {R3,R4,R8,R11,R17,R19,R20} |
| 24: {R3,R4,R7,R10,R17,R19,R20}  | 62: {R3,R4,R7,R11,R17,R19,R20} |
| 25: {R4,R8,R9,R10,R17,R18,R21}  | 63: {R3,R4,R10,R12,R17,R19}    |
| 26: {R4,R8,R9,R10,R14,R18,R21}  | 64: {R3,R4,R8,R10,R17,R19}     |
| 27: {R4,R8,R9,R10,R12,R18,R21}  | 65: {R3,R4,R10,R12,R17,R21}    |
| 28: {R4,R8,R9,R10,R11,R18,R21}  | 66: {R3,R4,R8,R10,R17,R21}     |
| 29: {R4,R6,R8,R9,R10,R18,R21}   | 67: {R3,R4,R7,R10,R17,R21}     |
| 30: {R4,R5,R8,R9,R10,R18,R21}   | 68: {R1,R9,R13,R18,R19,R20}    |
| 31: {R2,R4,R8,R9,R10,R18,R21}   | 69: {R1,R9,R11,R18,R19,R20}    |
| 32: {R4,R7,R9,R10,R17,R18,R21}  | 70: {R1,R9,R10,R18,R19,R20}    |
| 33: {R4,R7,R9,R10,R14,R18,R21}  | 71: {R1,R8,R9,R18,R19,R20}     |
| 34: {R4,R7,R9,R10,R12,R18,R21}  | 72: {R1,R7,R9,R18,R19,R20}     |
| 35: {R4,R7,R9,R10,R11,R18,R21}  | 73: {R1,R9,R13,R17,R19,R20}    |
| 36: {R4,R6,R7,R9,R10,R18,R21}   | 74: {R1,R9,R11,R17,R19,R20}    |
| 37: {R4,R5,R7,R9,R10,R18,R21}   | 75: {R1,R9,R10,R17,R19,R20}    |
| 38: {R2,R4,R7,R9,R10,R18,R21}   | 76: {R1,R8,R9,R17,R19,R20}     |

77:{R1,R7,R9,R17,R19,R20}	135:{R3,R4,R11,R17,R18,R20}
78:{R4,R8,R9,R10,R17,R18,R20}	136:{R3,R4,R11,R12,R17,R18}
79:{R4,R8,R9,R10,R14,R17,R18}	137:{R3,R4,R11,R12,R17,R19}
80:{R4,R8,R9,R10,R12,R17,R18}	138:{R1,R3,R4,R9,R17}
81:{R4,R8,R9,R10,R11,R17,R18}	139:{R2,R3,R4,R9,R17}
82:{R4,R6,R8,R9,R10,R17,R18}	140:{R3,R4,R9,R13,R17}
83:{R4,R7,R9,R10,R17,R18,R20}	141:{R3,R4,R9,R14,R17}
84:{R4,R7,R9,R10,R14,R17,R18}	142:{R3,R4,R9,R17,R19}
85:{R4,R7,R9,R10,R12,R17,R18}	143:{R3,R4,R9,R17,R21}
86:{R4,R7,R9,R10,R11,R17,R18}	144:{R1,R3,R4,R10,R17}
87:{R4,R6,R7,R9,R10,R17,R18}	145:{R2,R3,R4,R10,R17}
88:{R1,R4,R9,R10,R14,R18}	146:{R3,R4,R10,R13,R17}
89:{R1,R4,R9,R10,R13,R18}	147:{R3,R4,R10,R14,R17}
90:{R1,R4,R9,R10,R12,R18}	148:{R2,R4,R5,R8,R9,R17,R18,R20}
91:{R1,R4,R9,R10,R11,R18}	149:{R2,R4,R6,R8,R9,R17,R18,R20}
92:{R1,R4,R6,R9,R10,R18}	150:{R2,R4,R8,R9,R11,R17,R18,R20}
93:{R1,R9,R10,R14,R18,R19}	151:{R2,R4,R8,R9,R12,R17,R18,R20}
94:{R1,R9,R10,R12,R18,R19}	152:{R2,R4,R5,R8,R9,R13,R18,R20}
95:{R1,R9,R10,R11,R18,R19}	153:{R2,R4,R6,R8,R9,R13,R18,R20}
96:{R1,R6,R9,R10,R18,R19}	154:{R2,R4,R8,R9,R11,R13,R18,R20}
97:{R1,R9,R10,R14,R17,R19}	155:{R2,R4,R8,R9,R12,R13,R18,R20}
98:{R1,R9,R10,R13,R17,R19}	156:{R2,R4,R5,R7,R9,R17,R18,R20}
99:{R1,R9,R10,R12,R17,R19}	157:{R2,R4,R6,R7,R9,R17,R18,R20}
100:{R1,R9,R10,R11,R17,R19}	158:{R2,R4,R7,R9,R11,R17,R18,R20}
101:{R1,R6,R9,R10,R17,R19}	159:{R2,R4,R7,R9,R12,R17,R18,R20}
102:{R1,R4,R9,R10,R14,R17}	160:{R2,R4,R5,R7,R9,R13,R18,R20}
103:{R1,R4,R9,R10,R13,R17}	161:{R2,R4,R6,R7,R9,R13,R18,R20}
104:{R1,R4,R9,R10,R12,R17}	162:{R2,R4,R7,R9,R11,R13,R18,R20}
105:{R1,R4,R9,R10,R11,R17}	163:{R2,R4,R7,R9,R12,R13,R18,R20}
106:{R1,R4,R6,R9,R10,R17}	164:{R2,R4,R5,R8,R9,R18,R21}
107:{R1,R3,R4,R13,R17}	165:{R2,R4,R6,R8,R9,R18,R21}
108:{R2,R3,R4,R14,R17}	166:{R2,R4,R8,R9,R11,R18,R21}
109:{R2,R3,R4,R13,R17}	167:{R2,R4,R8,R9,R12,R18,R21}
110:{R3,R4,R11,R13,R17}	168:{R2,R4,R5,R7,R9,R18,R21}
111:{R1,R3,R5,R17,R19,R21}	169:{R2,R4,R6,R7,R9,R18,R21}
112:{R1,R3,R5,R17,R19,R20}	170:{R2,R4,R7,R9,R11,R18,R21}
113:{R1,R3,R5,R14,R17,R19}	171:{R2,R4,R7,R9,R12,R18,R21}
114:{R1,R3,R5,R13,R17,R19}	172:{R1,R3,R4,R17,R18,R21}
115:{R1,R3,R5,R10,R17,R19}	173:{R1,R3,R4,R7,R17,R20}
116:{R1,R3,R6,R17,R19,R21}	174:{R1,R3,R4,R8,R17,R20}
117:{R1,R3,R6,R17,R19,R20}	175:{R1,R3,R4,R11,R17,R20}
118:{R1,R3,R6,R14,R17,R19}	176:{R1,R3,R4,R17,R18,R20}
119:{R1,R3,R6,R13,R17,R19}	177:{R1,R3,R4,R17,R19,R20}
120:{R1,R3,R6,R10,R17,R19}	178:{R1,R3,R4,R7,R12,R17}
121:{R1,R3,R11,R17,R19,R20}	179:{R1,R3,R4,R8,R12,R17}
122:{R1,R3,R11,R14,R17,R19}	180:{R1,R3,R4,R11,R12,R17}
123:{R4,R8,R9,R17,R18,R19,R20}	181:{R1,R3,R4,R12,R17,R18}
124:{R4,R8,R9,R13,R18,R19,R20}	182:{R2,R3,R4,R7,R17,R21}
125:{R4,R7,R9,R17,R18,R19,R20}	183:{R2,R3,R4,R8,R17,R21}
126:{R4,R7,R9,R13,R18,R19,R20}	184:{R2,R3,R4,R7,R17,R20}
127:{R4,R8,R9,R10,R13,R18}	185:{R2,R3,R4,R8,R17,R20}
128:{R4,R7,R9,R10,R13,R18}	186:{R2,R3,R4,R11,R17,R20}
129:{R1,R4,R9,R18,R20}	187:{R2,R3,R4,R17,R18,R20}
130:{R1,R4,R9,R18,R21}	188:{R2,R3,R4,R17,R19,R20}
131:{R1,R9,R17,R19,R21}	189:{R2,R3,R4,R7,R12,R17}
132:{R1,R4,R9,R17,R20}	190:{R2,R3,R4,R8,R12,R17}
133:{R1,R4,R9,R17,R21}	191:{R2,R3,R4,R11,R12,R17}
134:{R1,R3,R12,R17,R19}	192:{R2,R3,R4,R12,R17,R18}

193:{R2,R3,R4,R12,R17,R19}	218:{R1,R5,R9,R10,R13,R18,R19}
194:{R1,R3,R4,R9,R18}	219:{R1,R7,R9,R10,R13,R18,R19}
195:{R3,R4,R6,R9,R18}	220:{R1,R8,R9,R10,R13,R18,R19}
196:{R3,R4,R7,R9,R18}	221:{R1,R3,R5,R9,R18,R19}
197:{R3,R4,R9,R12,R18}	222:{R1,R3,R6,R9,R18,R19}
198:{R3,R4,R9,R13,R18}	223:{R1,R3,R7,R9,R18,R19}
199:{R3,R4,R9,R17,R18}	224:{R1,R3,R8,R9,R18,R19}
200:{R3,R4,R9,R18,R19}	225:{R1,R3,R9,R12,R18,R19}
201:{R3,R4,R9,R18,R20}	226:{R1,R3,R9,R14,R18,R19}
202:{R3,R4,R9,R18,R21}	227:{R1,R3,R9,R18,R19,R20}
203:{R1,R3,R4,R5,R14,R17}	228:{R1,R5,R9,R18,R19,R21}
204:{R1,R3,R4,R6,R14,R17}	229:{R1,R6,R9,R18,R19,R21}
205:{R1,R3,R4,R12,R14,R17}	230:{R1,R7,R9,R18,R19,R21}
206:{R1,R3,R4,R14,R17,R18}	231:{R1,R8,R9,R18,R19,R21}
207:{R1,R3,R4,R14,R17,R20}	232:{R1,R9,R11,R18,R19,R21}
208:{R1,R3,R4,R5,R7,R17,R21}	233:{R1,R9,R12,R18,R19,R21}
209:{R1,R3,R4,R6,R7,R17,R21}	234:{R1,R9,R14,R18,R19,R21}
210:{R1,R3,R4,R5,R8,R17,R21}	235:{R1,R9,R18,R19,R20,R21}
211:{R1,R3,R4,R6,R8,R17,R21}	236:{R3,R4,R5,R10,R17,R18}
212:{R1,R3,R4,R11,R14,R17}	237:{R3,R4,R6,R10,R17,R18}
213:{R3,R4,R8,R11,R14,R17}	238:{R3,R4,R7,R10,R17,R18}
214:{R3,R4,R11,R14,R17,R18}	239:{R3,R4,R8,R10,R17,R18}
215:{R3,R4,R11,R14,R17,R19}	240:{R3,R4,R10,R11,R17,R18}
216:{R3,R4,R11,R14,R17,R20}	241:{R3,R4,R10,R12,R17,R18}
217:{R3,R4,R11,R14,R17,R21}	

La frecuencia con que los atributos aparecen en los reductos son las siguientes

Atributo	Nº Reductos	Frecuencia %
<b>R1</b>	109	45.23
<b>R2</b>	50	20.75
<b>R3</b>	126	52.28
<b>R4</b>	172	71.37
<b>R5</b>	25	10.37
<b>R6</b>	31	12.86
<b>R7</b>	57	23.65
<b>R8</b>	57	23.65
<b>R9</b>	146	60.58
<b>R10</b>	72	29.88



<b>R11</b>	54	22.41
<b>R12</b>	36	14.94
<b>R13</b>	35	14.52
<b>R14</b>	30	12.45
<b>R17</b>	152	63.07
<b>R18</b>	135	56.02
<b>R19</b>	104	43.15
<b>R20</b>	57	23.65
<b>R21</b>	78	32.37

La longitud de los reductos es de 8-5 atributos por reducto lo que indica que como mínimo, en este primer modelo, 11 atributos son redundantes, y en consecuencia podrían ser eliminados.

Los atributos que con más frecuencia aparecen en los reductos (al menos en un 40% de ellos) son, por orden de mayor frecuencia: R4, R17, R9, R18, R3, R1, R19.

En el segundo experimento hemos obtenido 229 reductos de la tabla de información codificada, y son:

- |                            |                            |
|----------------------------|----------------------------|
| 1: {R2,R3,R4,R7,R17,R21}   | 17: {R1,R3,R4,R17,R18,R20} |
| 2: {R2,R3,R4,R7,R17,R20}   | 18: {R1,R3,R4,R12,R17,R18} |
| 3: {R2,R3,R4,R7,R12,R17}   | 19: {R1,R3,R4,R6,R17,R18}  |
| 4: {R2,R3,R4,R6,R7,R17}    | 20: {R1,R3,R4,R5,R17,R18}  |
| 5: {R2,R3,R4,R5,R7,R17}    | 21: {R2,R3,R4,R9,R19,R21}  |
| 6: {R2,R3,R4,R17,R18,R21}  | 22: {R2,R3,R4,R9,R19,R20}  |
| 7: {R2,R3,R4,R17,R18,R20}  | 23: {R2,R3,R4,R9,R13,R19}  |
| 8: {R2,R3,R4,R12,R17,R18}  | 24: {R2,R3,R4,R9,R12,R19}  |
| 9: {R2,R3,R4,R6,R17,R18}   | 25: {R2,R3,R4,R9,R11,R19}  |
| 10: {R2,R3,R4,R5,R17,R18}  | 26: {R2,R3,R4,R6,R9,R19}   |
| 11: {R1,R3,R4,R7,R17,R21}  | 27: {R2,R3,R4,R5,R9,R19}   |
| 12: {R1,R3,R4,R7,R17,R20}  | 28: {R2,R3,R4,R9,R10,R21}  |
| 13: {R1,R3,R4,R7,R12,R17}  | 29: {R2,R3,R4,R9,R10,R20}  |
| 14: {R1,R3,R4,R6,R7,R17}   | 30: {R2,R3,R4,R9,R10,R13}  |
| 15: {R1,R3,R4,R5,R7,R17}   | 31: {R2,R3,R4,R9,R10,R12}  |
| 16: {R1,R3,R4,R17,R18,R21} | 32: {R2,R3,R4,R9,R10,R11}  |

33:	{R2,R3,R4,R6,R9,R10}	91:	{R1,R5,R8,R9,R10,R13,R19}
34:	{R2,R3,R4,R5,R9,R10}	92:	{R1,R2,R7,R9,R10,R13,R19}
35:	{R1,R3,R4,R9,R21}	93:	{R1,R5,R7,R9,R10,R13,R19}
36:	{R1,R3,R4,R9,R20}	94:	{R1,R3,R14,R17,R19}
37:	{R1,R3,R4,R9,R17}	95:	{R1,R3,R13,R17,R19}
38:	{R1,R3,R4,R9,R14}	96:	{R1,R3,R12,R17,R19}
39:	{R1,R3,R4,R9,R13}	97:	{R1,R3,R10,R17,R19}
40:	{R1,R3,R4,R9,R12}	98:	{R1,R5,R8,R9,R19,R21}
41:	{R1,R3,R4,R9,R11}	99:	{R1,R5,R7,R9,R19,R21}
42:	{R1,R3,R4,R8,R9}	100:	{R1,R2,R5,R9,R19,R21}
43:	{R1,R3,R4,R7,R9}	101:	{R1,R9,R18,R19,R20}
44:	{R1,R3,R4,R6,R9}	102:	{R1,R9,R17,R19,R20}
45:	{R1,R3,R4,R5,R9}	103:	{R1,R4,R9,R19,R20}
46:	{R1,R3,R8,R9,R17,R19}	104:	{R1,R3,R9,R19,R20}
47:	{R1,R3,R7,R9,R17,R19}	105:	{R1,R2,R9,R19,R20}
48:	{R1,R3,R6,R9,R17,R19}	106:	{R1,R9,R10,R17,R19}
49:	{R1,R3,R9,R11,R13,R18,R19}	107:	{R3,R4,R9,R14,R19}
50:	{R3,R4,R11,R14,R17,R21}	108:	{R3,R4,R9,R10,R14}
51:	{R3,R4,R11,R14,R17,R20}	109:	{R1,R9,R17,R19,R21}
52:	{R3,R4,R11,R14,R17,R19}	110:	{R1,R4,R9,R19,R21}
53:	{R3,R4,R8,R11,R14,R17}	111:	{R1,R3,R5,R9,R19}
54:	{R3,R4,R11,R12,R17,R19}	112:	{R4,R7,R9,R19,R21}
55:	{R3,R4,R8,R11,R17,R19}	113:	{R4,R7,R9,R19,R20}
56:	{R3,R4,R11,R12,R17,R21}	114:	{R4,R8,R9,R19,R21}
57:	{R3,R4,R8,R11,R17,R21}	115:	{R4,R8,R9,R19,R20}
58:	{R3,R4,R7,R11,R17,R21}	116:	{R1,R4,R9,R17,R21}
59:	{R3,R4,R10,R12,R17,R19}	117:	{R1,R4,R9,R18,R21}
60:	{R3,R4,R8,R10,R17,R19}	118:	{R1,R4,R9,R17,R20}
61:	{R3,R4,R10,R12,R17,R21}	119:	{R1,R4,R9,R18,R20}
62:	{R3,R4,R8,R10,R17,R21}	120:	{R2,R4,R7,R9,R17,R21}
63:	{R3,R4,R7,R10,R17,R21}	121:	{R2,R4,R7,R9,R18,R21}
64:	{R2,R4,R7,R9,R10,R20}	122:	{R2,R4,R7,R9,R17,R20}
65:	{R2,R4,R7,R9,R10,R17}	123:	{R2,R4,R7,R9,R18,R20}
66:	{R2,R4,R7,R9,R10,R12}	124:	{R2,R4,R8,R9,R17,R21}
67:	{R2,R4,R6,R7,R9,R10}	125:	{R2,R4,R8,R9,R18,R21}
68:	{R4,R7,R9,R10,R14,R20}	126:	{R2,R4,R8,R9,R17,R20}
69:	{R4,R7,R9,R10,R14,R17}	127:	{R2,R4,R8,R9,R18,R20}
70:	{R4,R7,R9,R10,R12,R14}	128:	{R3,R4,R9,R18}
71:	{R4,R6,R7,R9,R10,R14}	129:	{R3,R4,R11,R13,R17}
72:	{R4,R7,R9,R10,R17,R19}	130:	{R3,R4,R11,R17,R18}
73:	{R4,R7,R9,R10,R12,R19}	131:	{R3,R4,R10,R13,R17}
74:	{R4,R6,R7,R9,R10,R19}	132:	{R3,R4,R10,R14,R17}
75:	{R2,R4,R8,R9,R10,R20}	133:	{R4,R7,R9,R10,R13}
76:	{R2,R4,R8,R9,R10,R17}	134:	{R4,R7,R9,R10,R18}
77:	{R2,R4,R8,R9,R10,R12}	135:	{R4,R7,R9,R10,R21}
78:	{R2,R4,R6,R8,R9,R10}	136:	{R4,R8,R9,R10,R13}
79:	{R4,R8,R9,R10,R14,R20}	137:	{R4,R8,R9,R10,R18}
80:	{R4,R8,R9,R10,R14,R17}	138:	{R4,R8,R9,R10,R21}
81:	{R4,R8,R9,R10,R12,R14}	139:	{R2,R3,R4,R7,R9}
82:	{R4,R6,R8,R9,R10,R14}	140:	{R3,R4,R7,R9,R13}
83:	{R4,R8,R9,R10,R17,R19}	141:	{R3,R4,R7,R9,R14}
84:	{R4,R8,R9,R10,R12,R19}	142:	{R3,R4,R7,R9,R19}
85:	{R4,R6,R8,R9,R10,R19}	143:	{R3,R4,R7,R9,R21}
86:	{R1,R4,R7,R9,R10,R12}	144:	{R2,R3,R4,R8,R9}
87:	{R1,R4,R6,R7,R9,R10}	145:	{R3,R4,R8,R9,R13}
88:	{R1,R4,R8,R9,R10,R12}	146:	{R3,R4,R8,R9,R14}
89:	{R1,R4,R6,R8,R9,R10}	147:	{R3,R4,R8,R9,R19}
90:	{R1,R2,R8,R9,R10,R13,R19}	148:	{R3,R4,R8,R9,R21}

149:	{R3,R4,R9,R13,R17}	190:	{R1,R3,R7,R9,R10,R19}
150:	{R3,R4,R9,R14,R17}	191:	{R1,R3,R8,R9,R10,R19}
151:	{R3,R4,R9,R17,R19}	192:	{R1,R3,R9,R10,R11,R19}
152:	{R3,R4,R9,R17,R21}	193:	{R1,R3,R9,R10,R12,R19}
153:	{R1,R4,R9,R10,R17}	194:	{R1,R3,R9,R10,R14,R19}
154:	{R1,R4,R9,R10,R18}	195:	{R1,R5,R9,R18,R19,R21}
155:	{R1,R4,R9,R10,R20}	196:	{R1,R6,R9,R18,R19,R21}
156:	{R1,R4,R9,R10,R21}	197:	{R1,R7,R9,R18,R19,R21}
157:	{R1,R5,R8,R9,R19,R20}	198:	{R1,R8,R9,R18,R19,R21}
158:	{R1,R6,R8,R9,R19,R20}	199:	{R1,R9,R11,R18,R19,R21}
159:	{R1,R8,R9,R12,R19,R20}	200:	{R1,R9,R12,R18,R19,R21}
160:	{R1,R5,R7,R9,R19,R20}	201:	{R1,R9,R14,R18,R19,R21}
161:	{R1,R6,R7,R9,R19,R20}	202:	{R1,R3,R6,R9,R18,R19}
162:	{R1,R7,R9,R12,R19,R20}	203:	{R1,R3,R7,R9,R18,R19}
163:	{R1,R6,R8,R9,R10,R19}	204:	{R1,R3,R8,R9,R18,R19}
164:	{R1,R8,R9,R10,R12,R19}	205:	{R1,R3,R9,R12,R18,R19}
165:	{R1,R6,R7,R9,R10,R19}	206:	{R1,R3,R9,R14,R18,R19}
166:	{R1,R7,R9,R10,R12,R19}	207:	{R1,R2,R6,R9,R10,R19,R21}
167:	{R2,R3,R4,R8,R17}	208:	{R1,R2,R7,R9,R10,R19,R21}
168:	{R2,R3,R4,R9,R17}	209:	{R1,R2,R8,R9,R10,R19,R21}
169:	{R2,R3,R4,R10,R17}	210:	{R1,R2,R9,R10,R11,R19,R21}
170:	{R2,R3,R4,R11,R17}	211:	{R1,R2,R9,R10,R12,R19,R21}
171:	{R2,R3,R4,R13,R17}	212:	{R1,R2,R9,R10,R14,R19,R21}
172:	{R2,R3,R4,R14,R17}	213:	{R3,R4,R5,R10,R17,R18}
173:	{R2,R3,R4,R17,R19}	214:	{R3,R4,R6,R10,R17,R18}
174:	{R1,R3,R4,R8,R17}	215:	{R3,R4,R7,R10,R17,R18}
175:	{R1,R3,R4,R10,R17}	216:	{R3,R4,R8,R10,R17,R18}
176:	{R1,R3,R4,R11,R17}	217:	{R3,R4,R10,R12,R17,R18}
177:	{R1,R3,R4,R13,R17}	218:	{R1,R2,R3,R17,R19,R21}
178:	{R1,R3,R4,R14,R17}	219:	{R1,R3,R5,R17,R19,R21}
179:	{R1,R3,R4,R17,R19}	220:	{R1,R3,R6,R17,R19,R21}
180:	{R3,R4,R7,R11,R17,R19,R20}	221:	{R1,R3,R7,R17,R19,R21}
181:	{R3,R4,R7,R10,R17,R19,R20}	222:	{R1,R3,R8,R17,R19,R21}
182:	{R1,R5,R9,R10,R18,R19}	223:	{R1,R3,R11,R17,R19,R21}
183:	{R1,R6,R9,R10,R18,R19}	224:	{R1,R2,R3,R17,R19,R20}
184:	{R1,R7,R9,R10,R18,R19}	225:	{R1,R3,R5,R17,R19,R20}
185:	{R1,R8,R9,R10,R18,R19}	226:	{R1,R3,R6,R17,R19,R20}
186:	{R1,R9,R10,R11,R18,R19}	227:	{R1,R3,R7,R17,R19,R20}
187:	{R1,R9,R10,R12,R18,R19}	228:	{R1,R3,R8,R17,R19,R20}
188:	{R1,R9,R10,R14,R18,R19}	229:	{R1,R3,R11,R17,R19,R20}
189:	{R1,R3,R6,R9,R10,R19}		

La longitud de los reductos es de entre 4 y 7 atributos por reducto lo que indica que como mínimo, en este segundo modelo, 12 atributos son redundantes, y en consecuencia podrían ser eliminados.

Las frecuencias con que los atributos aparecen en los reductos son las siguientes:

<b>Atributo</b>	<b>Nº Reductos</b>	<b>Frecuencia %</b>
<b>R1</b>	116	50.66
<b>R2</b>	61	26.64
<b>R3</b>	133	58.08
<b>R4</b>	154	67.25
<b>R5</b>	20	8.73
<b>R6</b>	28	12.23
<b>R7</b>	59	25.76
<b>R8</b>	52	22.71
<b>R9</b>	156	68.12
<b>R10</b>	88	38.43
<b>R11</b>	24	10.48
<b>R12</b>	30	13.10
<b>R13</b>	18	7.86
<b>R14</b>	27	11.79
<b>R17</b>	98	42.79
<b>R18</b>	47	20.52
<b>R19</b>	107	46.72
<b>R20</b>	40	17.47
<b>R21</b>	52	22.71

Los atributos que aparecen con mayor frecuencia (más del 40%) en los reductos por orden de mayor frecuencia son: R9, R4, R3, R1, R19, R17.

De los resultados que hemos obtenido se pueden extraer dos primeras conclusiones.

La primera de ellas es que en ambos casos los ratios que con más frecuencia aparecen en los reductos son R1, R3, R4, R9, R17, R18 y R19 lo cual indica que son variables muy discriminatorias en la muestra entre empresas solventes y no solventes. Esto demuestra que a la hora de evaluar la solvencia de una empresa de seguros hay que tener en cuenta las siguientes cuestiones:

- a) La *liquidez* medida a través del ratio R1: Una de las cuestiones más importantes para asegurar el adecuado funcionamiento de cualquier empresa es la necesidad de tener suficiente liquidez. Pero en el caso de una empresa de seguros, este problema no debería aparecer, y si aparece revista una mayor importancia que en otro tipo de empresas, ya que la inversión del proceso productivo implica que las primas se cobran antes de que acaezcan los eventuales siniestros. Si una aseguradora no puede pagar los siniestros que ocurran, sus clientes y el público en general perderían la confianza en dicha compañía. Por otro lado, este ratio es una medida del equilibrio financiero si es positivo ya que indicaría que el capital circulante también lo es.
- b) La *rentabilidad* de las inversiones medida a través del ratio R3. Ya hemos comentado los rendimientos financieros de las entidades aseguradoras son en la actualidad la fuente principal de sus beneficios, o lo que es lo mismo, representan la posibilidad de obtener una rentabilidad adecuada para los recursos propios y también una fuente de financiación importante de los incrementos del Margen de Solvencia.
- c) Una política *apropiada* de *financiación* medida a través del ratio R4. Una financiación escasa debida a ingresos insuficientes que aparezcan de manera inesperada puede ser la causa del comienzo de proceso de quiebra.
- d) Unos adecuados *capitales propios* medidos a través del ratio R9. Se confirma la importancia, desde el punto de vista de la

solvencia, de estos fondos destinados a satisfacer, en caso necesario, futuros siniestros.

- e) La importancia de una conveniente *tarificación* medida a través de los ratios R17 y R18. Una mala tarificación lleva a una insuficiencia de las primas en relación con los costes tanto de los siniestros como del resto.
- f) Una *adecuada gestión del reaseguro* (R19).

El segundo comentario es que al menos 11 ó 12 atributos, según se trate de primer experimento o del segundo, son redundantes y podrían ser eliminados de la tabla. Esto demuestra la fuerza de esta metodología para la selección de las variables más significativas.

La intersección de los reductos nos da el núcleo. El núcleo de atributos en las dos pruebas realizadas estaba vacío. Esto indica que ningún único atributo es absolutamente necesario para la aproximación perfecta de las dos clases de decisión. Si el núcleo no hubiese sido vacío indicaría que existen atributos en el sistema que son indispensables desde el punto de vista discriminante, porque la no inclusión de cualquiera de los atributos contenidos en el núcleo conduce inmediatamente a una disminución de la calidad de la aproximación. Por otra parte, un núcleo no vacío ayuda a determinar los atributos más importantes además de las aproximación de las clases con las que está relacionado.

El siguiente paso consiste en la elección del reducto. En ambos casos para la selección del mismo hemos tenido en cuenta los siguientes criterios:

- a) El reducto debía contener el menor número de atributos como fuese posible.

b) Debía contener los atributos considerados, en nuestra opinión, más significativos para la evaluación de la solvencia de las empresas.

c) Después de seleccionar unos pocos reductos que cumplieren los dos primeros requisitos, la tercera condición es que el reducto que se fuese a seleccionar no tuviera ratios que estuviesen correlacionados. Por tanto hemos calculado algunos coeficientes de correlación.

Los reductos que hemos elegido para cada uno de las pruebas son el formado por **R1, R3, R9, R14, R17 y R19** para el primer modelo y el formado por **R3, R4, R9, R14, R17** para el segundo modelo.

Elegidos los reductos, el resto de atributos se pueden eliminar de la tabla de información codificada. Por tanto en el primer caso, con el reducto se pasó de un sistema inicial de información codificado de 19 columnas a otro de 6, y en el segundo caso se pasó a otro de 5 columnas.

### **II.3.- Resultados**

El sistema de información reducido puede verse como una tabla de decisión de la cual derivaremos las reglas de decisión.

De las posibles estrategias que incorpora ROSE para derivar las reglas de decisión de la tabla de información codificada y reducida, hemos seleccionado la estrategia consistente en la obtención del mínimo conjunto de reglas que cubren todos los objetos de la tabla de decisión.

El primer modelo obtenido se compone de 25 reglas y el segundo modelo obtenido se compone de 27 reglas. Las tablas 5.8. y 5.9 muestran los dos algoritmos obtenidos:



**TABLA 5.8. REGLAS DE DECISIÓN –MODELO 1: R1-R3-R9-R14-R17-R19**

Nº Regla	R1	R3	R9	R14	R17	R19	Decisión	Fuerza	Empresas
1	2	2					0	6	2,13,20,24,27,32
2				2	1		0	6	7,23,26,30,35,36
3		1	1				0	8	6,7,10,11,12,16,17,19
4	1		4				0	3	22,28,34
5		1	3				0	3	3,4,14
6					4	3	0	4	1,4,9,11
7			2			2	0	2	5,29
8		4		1			0	4	16,18,21,25
9	2				2	1	0	1	31
10	4		4		4		0	1	33
11	2			3			0	4	4,8,13,15
12		3		4			1	7	43,46,50,53,54,56,60
13			2		4		1	5	59,62,67,69,70
14	4			1			1	5	37,38,41,45,72
15	3	4	1				1	4	42,44,47,55
16	3				3		1	6	40,47,48,52,56,68
17	1		2			1	1	3	64,66,70
18	1		3			3	1	2	51,61
19			2	3			1	3	39,59,64
20		1		1		3	1	1	71
21	2	1		2			1	1	57
22	3		4		4		1	1	63
23	4		3				1	4	37,43,45,49
24				1	3		1	3	38,58,68
25	3			4			1	6	40,46,53,55,56,65

**TABLA 5.9. REGLAS DE DECISIÓN –MODELO 2: R3-R4-R9-R14-R17**

Nº Regla	R3	R4	R9	R14	R17	Decisión	Fuerza	Empresas
1	2		4			0	4	2, 14, 18, 23
2		3		2		0	4	17, 21, 25, 26
3	1	4				0	3	7, 10, 22
4		1	4			0	3	13, 19, 24
5			3	3		0	2	1, 4
6	3		1			0	2	6, 8
7	2		3	4		0	2	11, 15
8	4			1		0	2	12, 16
9	4	1				0	2	16, 20
10	2		1	3		0	1	9
11			3	2		0	1	3
12	3		2		2	0	1	5
13	2	2		2		0	3	2, 18, 27
14				1	3	1	3	29, 40, 50
15			2		4	1	5	41, 44, 49, 51, 52
16		1	3			1	4	28, 36, 38, 42
17	4	2				1	2	33, 37
18	3	4			1	1	2	30, 32
19	1	2	4			1	2	39, 47
20		2			2	1	1	53
21		3	1			1	1	31
22	2			1		1	2	43, 48
23	3			4		1	3	34, 38, 42
24	4			3		1	2	33, 35
25	1	3	2			1	1	46
26	1	1			2	1	1	54
27		4		2	4	1	1	45

Las reglas de decisión derivadas de la tabla de información codificada son, en ambos casos, reglas exactas dado que la calidad de la clasificación es igual a uno y, en consecuencia, las fronteras de las clases de decisión estaban vacías. Es decir, dado que los reductos dan una calidad de clasificación igual a la original, es decir a uno, todas las reglas

son determinísticas, esto es, describen únicamente la decisión que ha de tomarse cuando se satisfacen algunas condiciones.

En el primer modelo la fuerza de las reglas va de una longitud máxima de 7 a una mínima de 1 mientras que en el segundo modelo la longitud máxima es de 5 y la mínima vuelve a ser de 1.

Las reglas obtenidas constituyen para los dos modelos un algoritmo de clasificación. Son una representación no redundante del conocimiento contenido en la muestra seleccionada y están reunidas en un sistema de información codificado.

Si observamos en los algoritmos obtenidos solo se han utilizado 58 y 63 descriptores para cada modelo, lo que supone menos del 5% (4.24% y 4.6%, respectivamente) de los que aparecían en el sistema de información inicial (1.368).

En general, cuanto más corta es una regla más general es. En los dos conjuntos de reglas de decisión, el número de reglas de decisión que describen a las empresas no quebradas era mayor que el número de reglas que describían a las quebradas. Esto significa que, en nuestra muestra, cuesta más generalizar la descripción de las empresas sanas que las de las quebradas.

La muestra de 72 empresas utilizada en el estudio para derivar los algoritmos de decisión puede considerarse como una muestra de prueba utilizada para revelar las características financieras que discriminan las empresas solventes de las insolventes.

### III.- VALIDACIÓN DE LOS RESULTADOS: CLASIFICACIÓN DE EMPRESAS UTILIZANDO LOS DOS ALGORITMOS GENERADOS

Los dos algoritmos de clasificación pueden ser utilizados para evaluar cualquier otra empresa. Por tanto, la precisión predictiva debe contestarse en el curso de la verificación práctica ya que los algoritmos de clasificación obtenidos revelan únicamente qué criterios hay que tener en cuenta para evaluar la solvencia.

Por tanto las reglas de decisión obtenidas a partir del primer modelo han sido probadas con los datos de los años 2, 3, 4 y 5 antes del año base o año 1 (año antes de la quiebra). El segundo algoritmo obtenido ha sido validado con la muestra independiente formada por las 18 empresas que constituían la submuestra de validación.

Las siguientes tablas muestran los porcentajes de clasificaciones correctas obtenidas con la aplicación del método rough set para cada uno de los experimentos realizados:

**TABLA 5.10. MODELO 1**

<b>Clasificaciones correctas</b>	<b>Año 1</b>	<b>Año 2</b>	<b>Año 3</b>	<b>Año 4</b>	<b>Año 5</b>
Clase 0 (empresas quebradas)	100%	77.14%	55.56%	70.83%	63,16%
Clase 1 (empresas sanas)	100%	80%	76,67%	59,26%	76,19%
<b>Total</b>	100%	78.57%	66.67%	64,71%	70%

**TABLA 5.11. MODELO 2**

<b>Clasificaciones correctas</b>	<b>Año1</b>
Clase 0 (empresas quebradas)	77.78%
Clase 1 (empresas sanas)	77.78%
<b>Total</b>	77.78%

Debemos recordar que el programa ROSE utiliza una medida de distancia para reclasificar aquellas empresas que no encajan con ninguna regla y que es aplicada directamente, por lo que estos resultados ya incluyen aquellas empresas que no coincidiendo con ninguna regla son clasificadas con las reglas más cercanas según la medida de distancia ya comentada (ver capítulos 1 y 3).

En general y respecto a los dos modelos podemos comentar que los resultados obtenidos son bastante satisfactorios. Además, en el primer modelo, también se puede observar como lógicamente el porcentaje de clasificaciones correctas va disminuyendo conforme avanzamos hacia atrás en el tiempo y nos alejamos del año 1 o año previo a la quiebra excepto para el año 5. Una posible explicación puede ser que 5 años antes de la quiebra las empresas no tratan de manipular sus estados contables para ocultar los problemas financieros que atraviesan en vistas a que la situación puede ser reversible. Pero conforme esa situación, y tal y como hemos podido comprobar al analizar los balances y cuentas de resultados de las empresas que constituyen la muestra, se va complicando, empiezan a aparecer errores e irregularidades en las cuentas en aras de ocultar los problemas financieros por los que atraviesan. Pero llega un momento en que esa problemática situación es tan crítica que no se puede ocultar volviendo por tanto a mejorar en el año 2 la capacidad predictiva del modelo.

#### **IV.- COMPARACIÓN DEL ANÁLISIS ROUGH SET CON EL ANÁLISIS DISCRIMINANTE**

En general, la discriminación de tipo estadístico tiene gran aplicación cuando se trata de clasificar personas, entidades o cosas como pertenecientes a dos o más poblaciones dadas en base a ciertas observaciones sobre el valor que alcanzan unas variables, previamente elegidas, de carácter comúnmente cuantitativo.

En el caso de las finanzas aparecen aplicaciones obvias, por ejemplo:

- Clasificar, a los solicitantes de un crédito, en clientes potencialmente buenos o malos en función de determinados datos contables como pueden ser el grado de endeudamiento o la liquidez de su negocio.

- Clasificar, a un grupo de empresas, en candidatas potenciales al fracaso o al éxito en los negocios, según ciertos ratios contables tales como rentabilidad, endeudamiento, liquidez, proporción de beneficios retenidos, etc.

El análisis discriminante puede ser una buena herramienta de trabajo en el diagnóstico y pronóstico de crisis empresariales

Aunque la filosofía de los dos enfoques son muy diferentes, ambas metodologías se pueden aplicar a la predicción del fracaso empresarial sobre un conjunto de datos comunes.

El análisis discriminante es una técnica estadística que se utiliza para clasificar objetos en diferentes grupos basándose en la observación de algunas características de dichos objetos. Básicamente consiste en estimar una función lineal discriminante que calcula un resultado para cada objeto. Esta función es una combinación lineal ponderada de los valores para las características observadas del objeto. Las ponderaciones representan, en esencia, la importancia relativa y el impacto de las

características que forman parte de la función discriminante. El objeto se clasifica en un determinado grupo o clase basándose en la puntuación obtenida con la función discriminante.

El análisis discriminante está sujeto a un gran número de hipótesis tales como: cada grupo debe seguir una distribución normal multivariante; las matrices de covarianzas de cada grupo son idénticas; y, las matrices de covarianzas, las probabilidades a priori y los errores de clasificación han de ser conocidos. Si no se cumplen estas hipótesis los resultados que obtenemos pueden ser erróneos.

Sin embargo, en la práctica, estas hipótesis no se suelen verificar y muchos investigadores han cuestionado su aplicabilidad (Eisenbeis, 1977; Ohlson, 1980; Zavgren, 1985).

Las críticas a los modelos de análisis discriminante provienen de dos tipos de consideraciones. Las financiero-contables y las puramente estadísticas.

En cuanto a las críticas de tipo financiero una de las esgrimidas es la de tautología, en referencia a que las variables de ciertos modelos son idénticas o muy aproximadas a las definiciones de quiebra utilizadas para separar las empresas fallidas y no fallidas.

También parece razonable pensar que factores tales como la edad de la compañía, las consideraciones macroeconómicas y la variabilidad de los beneficios habrán de incorporarse en futuros análisis si se quieren desarrollar modelos verdaderamente predictivos.

Desde un punto de vista estadístico formal las principales críticas se han formulado por parte de Eisenbeis (1977) que señala los siguientes problemas:

- La distribución de las variables empleadas debería ser normal multivariada para que el análisis discriminante funcionase

correctamente. Sin embargo, en la práctica, los ratios contables no suelen presentar distribuciones normales.

- En el análisis de funciones lineales, como las empleadas por casi todos los modelos, se supone la igualdad de las matrices de covarianzas que indican el grado de dispersión de los grupos. Si dicha igualdad no se da, caso probable, habrían de emplearse mejor funciones cuadráticas.
- La significatividad de las variables del modelo no es fácil de dilucidar. No existen pruebas para detectar el impacto absoluto de una variable determinada. Solo pueden proponerse pruebas par determinar la importancia relativa de las variables individuales, tales como el test F a escala univariada y el ajuste de los coeficientes de la función discriminatoria al multiplicarlos por la diagonal de la matriz de covarianzas correspondiente.
- Cuando se reduce la dimensión del modelo eliminando ciertas variables dicha reducción afectará a los resultados de la clasificación, aunque se eliminen variables aparentemente no significativas.
- El análisis no puede aplicarse correctamente cuando una variable, intrínsecamente continua, se segmenta y se utiliza como una base de formación de grupos. Solo tiene sentido formar grupos, basados en una determinada variable, si aparece una ruptura o discontinuidad natural como en el caso de empresas en quiebra y no.
- Las reglas de clasificación deben incorporar las probabilidades “a priori” de fracaso y no fracaso y los costes relativos de los errores. En la práctica, estos aspectos no se tienen en cuenta.
- Si para estimar los errores del modelo se reclasifica la muestra original, la estimación conducirá a resultados altamente



sesgados y en exceso optimistas. Conviene, por ello, emplear muestras o submuestras diferentes, una para estimar la función y otra para evaluar los errores de clasificación.

Además debemos señalar como puntos débiles evidentes, en aplicaciones financieras del análisis discriminante, los siguientes:

- Limitaciones sectoriales o industriales. Los modelos funcionan mejor con muestras homogéneas de empresas en cuanto a sector o tipo de industria.
- Limitaciones dimensionales. Los predictores del fracaso de una gran empresa pueden no ser exactamente los mismos que en una pequeña, por ello suele acudir al emparejamiento de empresas, saneada y fracasada, por dimensión.
- Limitaciones geográficas. Los modelos parecen tener el sello específico de la nación donde han sido desarrollados. Las normas y estándares contables suelen variar de un país a otro y ello puede impedir la “exportación” de modelos.
- Limitaciones temporales. Los modelos no son perdurables en el tiempo. Los ciclos económicos en función del tiempo, pueden trastocar la aplicación de las funciones discriminantes.
- Aplicaciones a fenómenos extremos. La crisis observable ha de ser de carácter terminal, quiebra o liquidación, embotándose el análisis ante situaciones no extremas como problemas de liquidez o dificultades financieras pasajeras.
- Modelos no unívocos. Muchos estudios presentan una determinada función discriminante que es una entre muchas otras posibles, quizá solo con ligerísima ventaja predictiva sobre las demás. Ello suele deberse a la correlación existente entre los diferentes ratios utilizados.

Por lo tanto, en general , podemos hablar de falta de validación externa de los modelos expuestos aunque su validez interna suele ser bastante alta.

Como ventajas tendríamos:

- obtención de una puntuación única que permite una jerarquización de las empresas analizadas, así como una clasificación dicotómica entre saneadas y candidatas potenciales al fracaso.
- Utilización de los modelos, en el seno de una determinada empresa, para evaluar la propia situación financiera y poner en marcha las consiguientes acciones preventivas. Estaríamos ante el uso normativo de los modelos de predicción.

Por tanto, pese a estas últimas ventajas que hemos citado, en la práctica las hipótesis estadísticas que requiere el discriminante raramente se satisfacen. En consecuencia, en la mayoría de las ocasiones los ratios financieros no satisfacen los supuestos estadísticos mencionados por lo que cabría preguntarse por las razones que justifican su elección como método para comparar con la teoría rough set. La elección del discriminante se justifica porque ha sido el primer método que se aplicó y en consecuencia el más utilizado para problemas de clasificación y predicción dentro del campo financiero y de gestión, y porque ha dado muy buenos resultados empíricos en problemas financieros en los que intervienen como variables ratios contables pese a que estos no satisfagan las hipótesis estadísticas.

A diferencia del discriminante, el enfoque rough set no necesita que los datos verifiquen a prior ningún tipo de hipótesis pero una descripción más detallada y comparativa de los dos métodos se encuentra en Krusinnska et al. (1992) y Stefanowski (1992).

Para comparar los dos métodos hemos derivado una función discriminante utilizando los ratios del reducto seleccionado en su forma original (sin codificar). La siguiente tabla muestra los coeficientes obtenidos de la función discriminante:

<b>Variable</b>	<b>Coficiente</b>
Término Constante	-0,62358
R1	4,68606
R3	-0,29266
R9	0,24881
R14	-0,03074
R17	-0,76597
R19	1,86133

En cuanto a los porcentajes de clasificaciones correctas obtenidas con la aplicación del discriminante son las siguientes:(TABLA 5.12 )

**TABLA 5.12. RESULTADOS ANÁLISIS DISCRIMINANTE**

<b>Clasificaciones correctas</b>	<b>Año 1</b>	<b>Año 2</b>	<b>Año 3</b>	<b>Año 4</b>	<b>Año 5</b>
Clase 0 (empresas quebradas)	76.47%	78.79%	87.5%	72.72%	81.25%
Clase 1 (empresas sanas)	55.88%	64.71%	61.53%	43.47%	50%
<b>Total</b>	<b>66.17%</b>	<b>71.75%</b>	<b>74.51%</b>	<b>58.09%</b>	<b>65.62%</b>

Estos resultados provienen de asumir una determinada relación de probabilidades de clasificación errónea a la hora de estimar las probabilidades de pertenencia a cada grupo para cada unidad. Estas probabilidades estimadas de pertenencia a un grupo se utilizaron para clasificar las unidades en cada grupo, tomando como punto de corte una probabilidad de pertenencia a un grupo de 0.5, dado que en nuestro caso trabajamos con dos grupos. La idea es la siguiente normalmente se compensa el hecho de lo que te interesa captar (la insolvencia) tiene una probabilidad de ocurrencia reducida, con el hecho de que la valoración que se hace de cometer un error en el sentido de clasificar una unidad de

ese grupo reducido (insolventes) en el grupo más amplio (solventes) es mucho peor que el error ocurra en sentido contrario.

La siguiente tabla muestra los resultados comparativos de los dos enfoques, siempre en porcentajes de clasificaciones correctas:

**TABLA 5.13. RESULTADOS AMBOS MÉTODOS**

<b>Clasificaciones correctas</b>	<b>Año 1</b>	<b>Año 2</b>	<b>Año 3</b>	<b>Año 4</b>	<b>Año 5</b>
Rough Set	100%	78.57%	66.67%	64,71%	70%
Análisis Discriminante	66.17%	71.75%	74.51%	58.09%	65.62%

Tal y como puede verse en la tabla anterior los resultados del enfoque rough set, excepto para el año 3, son mejores que los del análisis discriminante. Estos resultados muestran la capacidad de esta metodología para responder de manera eficiente al problema de predicción del fracaso empresarial, siendo una alternativa muy fiable al análisis discriminante.

Es más, desde el punto de vista del usuario, las reglas de decisión expresan las dependencias entre las características financieras de una empresa y su situación en un lenguaje más accesible que una función discriminante. Una función discriminante solo da una visión global, mientras que las reglas de decisión hablan el lenguaje natural de las decisiones tomadas por un experto o decisor y que han servido como ejemplos.

## **ANEXO CAPITULO 5**

## Análisis Estadístico de los ratios

### Summary Statistics for R1

Count = 72  
Average = 0.270556  
Variance = 0.110941  
Standard deviation = 0.333077  
Minimum = -0.59  
Maximum = 0.98  
Range = 1.57  
Std. skewness = -1.50046  
Std. kurtosis = 0.694964

### Percentiles for R1

1.0% = -0.59  
5.0% = -0.32  
10.0% = -0.21  
25.0% = 0.115  
50.0% = 0.295  
75.0% = 0.475  
90.0% = 0.64  
95.0% = 0.85  
99.0% = 0.98

### Summary Statistics for R2

Count = 72  
Average = 0.0701389  
Variance = 0.0293056  
Standard deviation = 0.171189  
Minimum = -0.08  
Maximum = 1.2  
Range = 1.28  
Std. skewness = 16.1148  
Std. kurtosis = 47.0338

### Percentiles for R2

1.0% = -0.08  
5.0% = 0.0  
10.0% = 0.0  
25.0% = 0.0  
50.0% = 0.01  
75.0% = 0.07  
90.0% = 0.18  
95.0% = 0.35  
99.0% = 1.2

### Summary Statistics for R3

Count = 72  
Average = 0.0959722  
Variance = 0.0155934  
Standard deviation = 0.124874  
Minimum = 0.0  
Maximum = 0.73  
Range = 0.73  
Std. skewness = 11.7927  
Std. kurtosis = 23.8401

### Percentiles for R3

1.0% = 0.0  
5.0% = 0.0  
10.0% = 0.01  
25.0% = 0.03  
50.0% = 0.06  
75.0% = 0.11  
90.0% = 0.18  
95.0% = 0.33  
99.0% = 0.73

### Summary Statistics for R4

Count = 72  
Average = 0.155972  
Variance = 0.0459709  
Standard deviation = 0.214408  
Minimum = -0.37  
Maximum = 1.15  
Range = 1.52  
Std. skewness = 6.54386  
Std. kurtosis = 10.9518

### Percentiles for R4

1.0% = -0.37  
5.0% = -0.02  
10.0% = 0.0  
25.0% = 0.03  
50.0% = 0.08  
75.0% = 0.26  
90.0% = 0.38  
95.0% = 0.57  
99.0% = 1.15

Summary Statistics for R5

Count = 72  
Average = 20.4122  
Variance = 24907.0  
Standard deviation = 157.819  
Minimum = -10.91  
Maximum = 1340.66  
Range = 1351.57  
Std. skewness = 29.3736  
Std. kurtosis = 124.592

Percentiles for R5

1.0% = -10.91  
5.0% = -4.62  
10.0% = 0.06  
25.0% = 0.565  
50.0% = 1.565  
75.0% = 3.29  
90.0% = 6.6  
95.0% = 7.58  
99.0% = 1340.66

Summary Statistics for R6

Count = 72  
Average = 19.3869  
Variance = 22927.3  
Standard deviation = 151.418  
Minimum = -10.91  
Maximum = 1286.11  
Range = 1297.02  
Std. skewness = 29.3758  
Std. kurtosis = 124.604

Percentiles for R6

1.0% = -10.91  
5.0% = -4.28  
10.0% = 0.06  
25.0% = 0.525  
50.0% = 1.38  
75.0% = 2.715  
90.0% = 4.94  
95.0% = 6.98  
99.0% = 1286.11

Summary Statistics for R7

Count = 72  
Average = 1.86431  
Variance = 40.5553  
Standard deviation = 6.3683  
Minimum = 0.01  
Maximum = 53.7  
Range = 53.69  
Std. skewness = 27.1696  
Std. kurtosis = 111.189

Percentiles for R7

1.0% = 0.01  
5.0% = 0.09  
10.0% = 0.29  
25.0% = 0.455  
50.0% = 0.725  
75.0% = 1.22  
90.0% = 2.7  
95.0% = 3.74  
99.0% = 53.7

Summary Statistics for R8

Count = 72  
Average = 1.65278  
Variance = 40.3128  
Standard deviation = 6.34924  
Minimum = -5.26  
Maximum = 53.7  
Range = 58.96  
Std. skewness = 27.5966  
Std. kurtosis = 114.5

Percentiles for R8

1.0% = -5.26  
5.0% = 0.05  
10.0% = 0.22  
25.0% = 0.46  
50.0% = 0.7  
75.0% = 1.18  
90.0% = 1.9  
95.0% = 3.13  
99.0% = 53.7

Summary Statistics for R9

Count = 72  
 Average = 0.389722  
 Variance = 0.161465  
 Standard deviation = 0.401827  
 Minimum = -1.16  
 Maximum = 1.0  
 Range = 2.16  
 Std. skewness = -3.77911  
 Std. kurtosis = 5.27407

Percentiles for R9

1.0% = -1.16  
 5.0% = -0.16  
 10.0% = 0.06  
 25.0% = 0.14  
 50.0% = 0.35  
 75.0% = 0.68  
 90.0% = 0.91  
 95.0% = 0.96  
 99.0% = 1.0

Summary Statistics for R10

Count = 72  
 Average = 24.0114  
 Variance = 36804.4  
 Standard deviation = 191.845  
 Minimum = -11.17  
 Maximum = 1629.05  
 Range = 1640.22  
 Std. skewness = 29.3821  
 Std. kurtosis = 124.64

Percentiles for R10

1.0% = -11.17  
 5.0% = -1.66  
 10.0% = 0.0  
 25.0% = 0.04  
 50.0% = 0.545  
 75.0% = 2.97  
 90.0% = 6.04  
 95.0% = 6.67  
 99.0% = 1629.05

Summary Statistics for R11

Count = 72  
 Average = 18.1181  
 Variance = 20493.5  
 Standard deviation = 143.155  
 Minimum = -7.81  
 Maximum = 1215.8  
 Range = 1223.61  
 Std. skewness = 29.3817  
 Std. kurtosis = 124.638

Percentiles for R11

1.0% = -7.81  
 5.0% = -3.01  
 10.0% = 0.0  
 25.0% = 0.27  
 50.0% = 1.095  
 75.0% = 2.43  
 90.0% = 4.94  
 95.0% = 5.48  
 99.0% = 1215.8

Summary Statistics for R12

Count = 72  
 Average = 17.6903  
 Variance = 20132.4  
 Standard deviation = 141.889  
 Minimum = -6.89  
 Maximum = 1204.81  
 Range = 1211.7  
 Std. skewness = 29.3843  
 Std. kurtosis = 124.653

Percentiles for R12

1.0% = -6.89  
 5.0% = -2.8  
 10.0% = 0.0  
 25.0% = 0.27  
 50.0% = 0.845  
 75.0% = 1.815  
 90.0% = 2.93  
 95.0% = 4.54  
 99.0% = 1204.81



Summary Statistics for R13

Count = 72  
Average = 1.22292  
Variance = 16.238  
Standard deviation = 4.02964  
Minimum = 0.0  
Maximum = 33.65  
Range = 33.65  
Std. skewness = 26.4092  
Std. kurtosis = 106.154

Percentiles for R13

1.0% = 0.0  
5.0% = 0.01  
10.0% = 0.14  
25.0% = 0.27  
50.0% = 0.49  
75.0% = 0.82  
90.0% = 1.57  
95.0% = 2.23  
99.0% = 33.65

Summary Statistics for R14

Count = 72  
Average = 1.03278  
Variance = 15.993  
Standard deviation = 3.99912  
Minimum = -4.46  
Maximum = 33.65  
Range = 38.11  
Std. skewness = 27.1092  
Std. kurtosis = 112.157

Percentiles for R14

1.0% = -4.46  
5.0% = 0.0  
10.0% = 0.1  
25.0% = 0.225  
50.0% = 0.435  
75.0% = 0.765  
90.0% = 1.54  
95.0% = 2.02  
99.0% = 33.65

Summary Statistics for R17

Count = 72  
Average = 1.25361  
Variance = 1.03725  
Standard deviation = 1.01846  
Minimum = 0.66  
Maximum = 9.5  
Range = 8.84  
Std. skewness = 26.6412  
Std. kurtosis = 108.716

Percentiles for R17

1.0% = 0.66  
5.0% = 0.87  
10.0% = 0.91  
25.0% = 0.98  
50.0% = 1.055  
75.0% = 1.27  
90.0% = 1.5  
95.0% = 1.74  
99.0% = 9.5

Summary Statistics for R18

Count = 72  
Average = 1.30111  
Variance = 1.04942  
Standard deviation = 1.02441  
Minimum = 0.73  
Maximum = 9.5  
Range = 8.77  
Std. skewness = 25.7814  
Std. kurtosis = 103.575

Percentiles for R18

1.0% = 0.73  
5.0% = 0.92  
10.0% = 0.94  
25.0% = 1.0  
50.0% = 1.09  
75.0% = 1.29  
90.0% = 1.57  
95.0% = 1.74  
99.0% = 9.5

Summary Statistics for R19

Count = 68  
Average = 0.148529  
Variance = 0.0453172  
Standard deviation = 0.212878  
Minimum = 0.0  
Maximum = 0.89  
Range = 0.89  
Std. skewness = 5.91814  
Std. kurtosis = 4.46797

Percentiles for R19

1.0% = 0.0  
5.0% = 0.0  
10.0% = 0.0  
25.0% = 0.0  
50.0% = 0.065  
75.0% = 0.19  
90.0% = 0.48  
95.0% = 0.54  
99.0% = 0.89

Summary Statistics for R20

Count = 72  
Average = 0.643889  
Variance = 0.0664495  
Standard deviation = 0.257778  
Minimum = 0.0  
Maximum = 1.22  
Range = 1.22  
Std. skewness = -2.15329  
Std. kurtosis = 1.01208

Percentiles for R20

1.0% = 0.0  
5.0% = 0.06  
10.0% = 0.29  
25.0% = 0.515  
50.0% = 0.68  
75.0% = 0.785  
90.0% = 0.95  
95.0% = 1.05  
99.0% = 1.22

Summary Statistics for R21

Count = 72  
Average = 0.619306  
Variance = 0.0671305  
Standard deviation = 0.259096  
Minimum = -0.01  
Maximum = 1.38  
Range = 1.39  
Std. skewness = -1.46278  
Std. kurtosis = 1.93368

Percentiles for R21

1.0% = -0.01  
5.0% = 0.02  
10.0% = 0.28  
25.0% = 0.515  
50.0% = 0.655  
75.0% = 0.75  
90.0% = 0.94  
95.0% = 0.99  
99.0% = 1.38

## **CONCLUSIONES**

Ha llegado el momento de ver en qué medida hemos dado una respuesta adecuada a las tres cuestiones que han motivado nuestro trabajo.

La primera de ellas trataba de explicar por qué, de entre los problemas financieros que podíamos acometer elegíamos la **predicción de crisis empresariales**. La respuesta era clara, y es que a diferencia de otros problemas financieros, el gran número de agentes e intereses afectados cuando se produce una insolvencia, hacen que el estudio de la misma se mantenga siempre actual. Es más, la viabilidad y, en consecuencia, el riesgo de fracaso empresarial ha sido reconocido como un factor importante en el área de la macroeconomía o en los análisis industriales. Al mismo tiempo, tanto el número de empresas quebradas como la tendencia relativa, son considerados indicadores de la solidez de las industrias y del conjunto de la economía.

Al hilo de esta primera cuestión argumentábamos el por qué la insolvencia, su temprana detección o el conocimiento de las condiciones que pueden conducir a ella, en una compañía de **seguros** es una de las principales preocupaciones de los legisladores, consumidores y directivos de este tipo de entidades. Esta preocupación surgía de la necesidad de proteger al público de las consecuencias de las insolvencias de los aseguradores y la necesidad de minimizar la carga que supone para el estado hacer frente a las mismas a través de los fondos de garantía. Como consecuencia, veíamos cómo una de las funciones más importantes de los gobiernos es regular el sector asegurador y a través de ella, controlar la solvencia del mismo.

En definitiva, comprobamos cómo el análisis de la solvencia en cualquier sector es importante, pero la creciente importancia de las entidades aseguradoras en cualquier sistema financiero, hacen que dicha cuestión sea clave en este sector.

Nuestra investigación recoge como **novedad** respecto a otros estudios empíricos que aplican la metodología Rough Set a la predicción del fracaso empresarial, el desarrollo de unos modelos para las entidades aseguradoras no-vida, y su aplicación al caso español.

La última pregunta que nos formulábamos, cuestionaba el por qué de la elección de la **Teoría Rough Set** como método a aplicar a la predicción del fracaso empresarial en el sector seguros no-vida. Cabría recordar que es un enfoque que se encuadra dentro de las aplicaciones de la Inteligencia Artificial. Utiliza la experiencia de una forma objetiva, a través del estudio de la experiencia histórica de una manera cuantitativa, para así explicitar reglas que, resumiendo y objetivando esa experiencia acumulada, ayuden en las decisiones futuras.

Es un método originalmente desarrollado en los años ochenta por **Pawlak**, entre otros, como herramienta formal para tratar con la **incertidumbre** o la vaguedad inherente a un proceso de decisión. Se podría decir que existen tres categorías generales de imprecisión en el análisis científico. La primera categoría ocurre, cuando un acontecimiento es aleatorio por naturaleza y la imprecisión asociada con este tipo de acontecimientos puede describirse por la teoría estadística de la probabilidad. La segunda surge del hecho de que los objetos pueden no pertenecer exclusivamente a una única categoría sino que pueden pertenecer a varias categorías aunque con diferentes grados; en este caso la imprecisión toma forma de pertenencia difusa a un conjunto y es el objeto de la lógica *fuzzy*. La tercera categoría es la teoría rough set que es útil cuando las clases en las que han de clasificarse los objetos son imprecisas, pero sin embargo pueden aproximarse mediante conjuntos precisos (Nurmi et al., 1996).

Cuando se trabaja con compañías reales, el análisis de las relaciones causa-efecto relativas a la predicción de insolvencias, se compone normalmente de inconsistencias en las clasificaciones. Es decir, una empresa sana puede tener los mismos atributos que una quebrada.

De entre las tres teorías que abarcan el análisis de la imprecisión, un enfoque muy adecuado para tratar esta problemática es el Rough Set.

Además este método, está especialmente concebido para problemas de **clasificación multiatributo**. La predicción del fracaso empresarial es un ejemplo claro de este tipo de problemas ya que pretendemos asignar empresas (nuestros objetos), descritas por los valores de un conjunto de ratios financieros (nuestros atributos), a una clase o categoría (sana o fracasada).

Por otro lado, en comparación con otros métodos, el enfoque rough set ofrece las siguientes **ventajas**:

- No necesita ninguna información adicional como probabilidades (en estadística) o grado de pertenencia (en la teoría fuzzy set). Es conceptualmente simple y necesita algoritmos sencillos.

- Reduce los atributos y objetos superfluos para obtener los denominados *reductos* (o *conjuntos mínimos de atributos*) que aseguran la misma calidad de clasificación que el conjunto total. A través de ellos se obtiene la mínima representación del conocimiento en forma de reglas de decisión teniendo en consecuencia, una interpretación sencilla. Es decir, el modelo que obtenemos, o más precisamente, la predicción del modelo, es en forma de reglas de decisión y éstas tienen en cuenta las preferencias del decisor que toma parte en el proceso de construcción de las mismas. Es más, las reglas de decisión revelan los atributos más relevantes que debería considerar el decisor para evaluar el riesgo de fracaso de una empresa.

El modelo de preferencia global en forma de reglas de decisión derivadas de un conjunto de ejemplos o de casos tiene ventajas sobre un modelo funcional o relacional ya que explica las preferencias a través de los hechos importantes en términos únicamente de los

criterios más significativos y de manera fácilmente comprensible. Las reglas se justifican a través de los objetos o ejemplos de los que se derivan y, aún más, las inconsistencias que se manifiestan en los ejemplos ni se corrigen ni se agregan por una función o relación global. Esto se debe a que está especialmente indicada para tratar las inconsistencias que se presentan en los ejemplos o en los casos; no corrige dichas inconsistencias sino que calcula aproximaciones "por arriba" y "por abajo". En consecuencia, y basándose en estas aproximaciones, se inducen dos conjuntos de reglas: determinísticas (ciertas) y no-determinísticas (posibles). El conjunto de reglas se puede utilizar para justificar las decisiones relativas a un nuevo objeto. Esta justificación se logra mediante el emparejamiento del nuevo objeto a una de las reglas; si no es posible, se presenta al decisor el conjunto de reglas "más cercanas". Para ello se han presentado dos medidas de distancia, la métrica  $L_p$  y la relación de proximidad valorada.

Por tanto, descubre los hechos importantes escondidos en los datos y lo expresa en el lenguaje natural de las reglas de decisión.

- Acepta tanto atributos cualitativos como cuantitativos (incluyendo aquellos cuyos dominios no están ordenados) y especifica su importancia mediante la aproximación de la clasificación. Además los modelos clásicos para la predicción del fracaso empresarial (por ejemplo modelos basados en el análisis discriminante) dan únicamente una medida financiera del riesgo, ignorando otras medidas importantes del mismo como son la experiencia de los directivos, el nicho de mercado o las ventajas competitivas que posea la empresa. Es decir, el modelo construido a través de reglas parece más cercano a una medida real del riesgo potencial de quiebra al tener en consideración variables de tipo cualitativo, que aunque nosotros en nuestro estudio no hemos podido introducir por carecer de información relativa a los mismos, hemos comprobado en otros trabajos su posible

introducción, y que conduciría junto con los ratios a una evaluación global del riesgo.

- Contribuye a la minimización de tiempo y de coste del proceso de toma de decisión (el enfoque rough set es un sistema que procesa la información en tiempo real).

- Da transparencia a las decisiones de clasificación, lo que permite su discusión y refrendo.

- Toma en consideración todo el conocimiento adquirido del decisor.

- Finalmente, la utilización de un método multicriterio para abordar decisiones financieras en tiempo real, implica la necesidad de disponer de un software que permita la aplicación práctica de dicho método. **ROSE**, es el que hemos utilizado en la parte empírica de nuestra investigación. Por tanto, en el ámbito económico-financiero no se ha empleado este nuevo software para abordar un problema financiero real como es la predicción del fracaso empresarial. En consecuencia, esta sería una segunda **aportación original** de nuestra investigación.

Las expectativas que originariamente motivaron el inicio de nuestro trabajo, en el que hemos tratado de comprobar empíricamente la utilidad e idoneidad del método Rough Set para la predicción de crisis empresariales en las empresas de seguros, mediante el desarrollo de **un modelo** que ha sido aplicado a empresas aseguradoras españolas en el ramo no-vida y, que incorpora como factores de dicho análisis ratios financieros, creemos que han dado unos resultados, incluso superiores a los inicialmente esperados.

En el desarrollo de nuestra investigación, hemos partido de una información contable (Balances y Cuentas de Resultados) relativa a una muestra de empresas españolas de seguros no vida y hemos obtenido un



conjunto de ratios financieros indicativos de solvencia (directa o indirectamente). Estos **ratios** calculados serán las *variables explicativas* del riesgo de insolvencia y aunque es una información demasiado agregada presenta la ventaja de ser obligatoria, homogénea y pública. Al hilo de esta cuestión queremos insistir en la imposibilidad de haber introducido variables de tipo cualitativo u otras informaciones como el margen de solvencia (aspectos que hubiesen enriquecido de gran manera nuestros modelos), pero la imposibilidad de conseguirlas en el primer caso, y la no publicidad del mismo y de otros datos técnicos en el periodo analizado, en el segundo caso, han imposibilitado su inclusión.

Conviene recordar que, la elección de las variables más adecuadas a utilizar en la elaboración de un modelo de predicción es parte fundamental del éxito final del mismo. En nuestro caso, hemos configurado un conjunto de variables, que, a nuestro juicio, y, desde un punto de vista teórico y práctico, forman un modelo integral de análisis financiero de una entidad aseguradora, y que consideramos que son las más relevantes en relación con la solvencia. Para ello hemos acudido a los más utilizados en el análisis financiero para el pronóstico de la insolvencia de entidades aseguradoras (tanto generales como específicos), teniendo en cuenta nuestro criterio personal, dadas las propias características del sector seguros, y teniendo en cuenta un análisis preliminar de las características de los dos grupos de empresas (sanas y fracasadas).

Ya anticipamos, que en nuestro estudio no íbamos a partir de una gran batería de ratios porque introducir muchas variables en un modelo puede acarrear el efecto contrario al que buscamos: en vez de incrementar la capacidad predictiva, estamos introduciendo “ruido” en el mismo que puede disminuir dicha capacidad.

Es necesario resaltar, que a lo largo de la literatura contable se han propuesto muchos ratios para la medición de la insolvencia, tanto generales como específicos para el sector asegurador, y por tanto, otros

podrían haberse utilizado, pero el objeto de este trabajo es mostrar la utilidad de esta metodología en la predicción de las crisis empresariales, la flexibilidad del método utilizado permite que las variables puedan ser cualitativas o cuantitativas, o que los ratios sean unos u otros o cambiarlos en función de las condiciones económicas o del sector que se considere, o que el número de ellos sea mayor o menor.

Efectuamos un breve análisis descriptivo de los ratios que ha mostrado como, para la mayoría de ellos, las empresas sanas presentan mejores valores que las empresas fracasadas.

Obtenido el sistema de información inicial (72 empresas descritas mediante 19 ratios) realizamos dos tipos de experimentos, obteniendo de cada uno de ellos dos modelos (dos conjuntos de reglas de decisión), por lo que nos referiremos a ellos como **modelo 1** y **modelo 2**.

Las reglas de decisión del *modelo 1* se obtuvieron empleando los datos del año 1 (año anterior a la quiebra) y verificando la capacidad predictiva de las mismas utilizando los datos de los años 2, 3, 4 y 5 previos a la quiebra. Posteriormente estimamos una **función discriminante** con el año 1 y comprobamos su capacidad predictiva utilizando los años 2, 3, 4 y 5 anteriores a que las firmas quebraran, al objeto de comparar ambos métodos.

Dado que utilizar las mismas empresas para validar y para estimar podía condicionar los resultados obtenidos, obtuvimos un *segundo modelo* seleccionado para el año 1 (año previo a la quiebra), 27 empresas sanas y 27 fracasadas de manera aleatoria para estimar las reglas de decisión, dejando las 18 restantes (9 sanas y 9 fracasadas) para comprobar la validez de las mismas. Pretendíamos de esta forma aplicar el modelo a datos que no hubieran sido utilizados para la elaboración del mismo.

Mediante el empleo de la metodología rough set obtuvimos las variables más significativas. Elegidos los reductos, el resto de atributos se pueden eliminar de la tabla de información codificada. Por tanto en el primer caso, con el reducto se pasó de un sistema inicial de información codificado de 19 columnas a otro de 6, y en el segundo caso se pasó a otro de 5 columnas. Los dos sistemas reducidos proporcionan la misma calidad de clasificación que la totalidad del conjunto de atributos. Demostramos, así, la fuerza de este enfoque en cuanto a selección de las características o variables más significativas de un sistema de información. Esto permite centrarse en aquellas variables clave reduciendo el tiempo y el coste del proceso de decisión.

En este estudio, el análisis rough set ha recalcado la importancia de los ratios financieros de liquidez, rentabilidad, cashflow, capacidad de endeudamiento, además de una adecuada tarificación, suficiencia de las provisiones y una conveniente política de reaseguro.

A partir de los reductos, obtuvimos los dos modelos consistentes en dos algoritmos compuestos, el primero de ellos por 25 reglas y el segundo de ellos por 27 reglas.

Las **reglas** obtenidas constituyen para los dos modelos un algoritmo de clasificación. Son una representación no redundante del conocimiento contenido en la muestra seleccionada y están reunidas en un sistema de información codificado. Es decir, la muestra de 72 empresas utilizada en el estudio para derivar los algoritmos de decisión puede considerarse como una muestra de prueba empleada para revelar las características financieras que discriminan las empresas solventes de las insolventes.

Si observamos en los algoritmos obtenidos solo se han utilizado 58 y 63 descriptores para cada modelo, lo que supone menos del 5% (4.24% y 4.6%, respectivamente) de los que aparecían en el sistema de información inicial (1.368).

Los algoritmos de clasificación obtenidos únicamente muestran qué criterios hay que tener en cuenta para evaluar la solvencia, pero es necesario validarlos de acuerdo con lo expuesto anteriormente para cada modelo. Es decir, los reductos obtienen el más pequeño subconjunto de ratios con el mayor poder discriminante en relación con el fracaso empresarial, y a través de esos subconjuntos se elaboran los modelos, las reglas. Estas muestran la relación entre las variables financieras y la solvencia o insolvencia de las empresas estudiadas.

Los **resultados** del primer modelo, en términos de clasificaciones correctas son 100%, 78.57%, 66.67%, 64,71% y 70% para los años 1, 2, 3, 4 y 5. Los obtenidos con la función discriminante estimada son 66.17%, 71.75%, 74.51%, 58.09% y 65.62%. Los resultados obtenidos con la metodología Rough Set son bastante satisfactorios y, en general son superiores, en nuestra muestra, a los obtenidos con el análisis discriminante.

Esta conclusión es muy importante porque convierte al enfoque rough set en el futuro, en una fuerte alternativa para el análisis de los problemas de gestión financiera.

Para el segundo modelo obtuvimos un porcentaje de aciertos del 77.78%, lo cual es un resultado bastante bueno ya que el modelo se ha validado sobre una muestra independiente.

Es necesario mencionar que las reglas se han derivado de unos datos particulares y representan una descripción general de la experiencia de un decisor en particular. De acuerdo con esto, las reglas no pueden aplicarse de manera general. Para aplicar a otros sectores, deberían utilizarse nuevos datos, analizarlos y generar las reglas adecuadas, que representarán la experiencia del nuevo decisor.

Por otro lado, el estudio tiene una serie de **limitaciones**. Una de ellas es que las compañías seleccionadas eran sociedades anónimas con

lo que se han excluido otras formas societarias del sector seguros y, en consecuencia, no es representativo de las mismas. Los modelos se componen de 25 y 27 reglas lo cual los hace relativamente complejos, aunque no lo son tanto teniendo en cuenta que los datos originales se componían de 19 ratios. Cabe suponer que si las variables son menores (como así lo demuestran otras aplicaciones del Rough Set al problema de predicción del fracaso empresarial) el número de reglas resultantes sería mucho menor lo que contribuiría aún más a la simplicidad en el uso del modelo. Otra limitación es que los resultados del modelo asumen igualdad de costes para los errores de clasificación tanto para las fracasadas como para las sanas. Estos errores son los denominados error de Tipo I y el error de Tipo II. Esta suposición es en realidad poco sostenible cuando nos enfrentamos a problemas reales ya que los decisores no valoran igualmente estos dos tipos de errores en la clasificación.

En cuanto a la verificación de los modelos, es evidente que no podemos obtener conclusiones significativas dado el escaso número de empresas analizadas, pero podemos observar que el nivel de aciertos es muy superior al que cabría esperar de un modelo aleatorio y que existe un mayor nivel de aciertos cuanto más próximo está el momento de la crisis.

Como **aplicaciones** prácticas, esta metodología puede ser muy útil para establecer un “sistema de alerta temprana” de detección de insolvencia para entidades aseguradoras ya que los reductos nos permiten centrarnos en un conjunto mínimo de variables. Este sistema podría ser muy útil para las autoridades reguladoras de este tipo de empresas, inversores, directivos, analistas financieros, bancos, auditores, asegurados y consumidores ya que centrándose en muy pocos ratios pueden controlar y revisar gran cantidad de firmas reduciendo por tanto el tiempo y los costes empleados en el proceso.

En definitiva, pese a las limitaciones mencionadas y la falta de generalización derivada del escaso número de empresas analizadas, nuestra investigación confirma los resultados de investigaciones previas

que han utilizado el método rough set demostrando que este enfoque proporciona un modelo para la predicción del fracaso empresarial con un alto porcentaje de aciertos y fácilmente interpretable. Por tanto, a los investigadores y decisores del sector asegurador, les resultaría muy útil tener en cuenta esta teoría en relación con la detección precoz de la insolvencia empresarial o en relación a otros problemas financieros.

## **BIBLIOGRAFÍA**

**Aguirre, M.** (coord.) (1987). *Biblioteca práctica de contabilidad*, vol. 6, *Contabilidades especiales*.

**Altman, E.I.** (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *The Journal of Finance* 23, 589-609.

**Altman, E. I.** (1993). *Corporate Financial Distress and Bankruptcy*, John Wiley and Sons, New York.

**Altman, E.I., Avery, R., Eisebenis, R., y Stinkey, J.** (1981). *Application of classification techniques in business, banking and finance*, Contemporary Studies in Economic and Financial Analysis, vol. 3. JAI Press, Greenwich.

**Altman, E. I., Marco, G. y Varetto, F.** (1994). Corporate distress diagnosis: Comparisons using discriminant analysis and neural networks (the Italian experience), *Journal of Banking and Finance* 18, 505-529.

**Álvarez, S. y Urías, J.** (1987). *Análisis Contable Superior*, Universidad Nacional de Educación a Distancia.

**Ambrose, J.M. y Carroll, A.M.** (1994). Using Best's Ratings in Life Insurer Insolvency Prediction, *The Journal of Risk and Insurance* Vol. 61, 2, 317-327.

**Ambrose, J.M., y Seward, J.A.** (1988). Best's Ratings, Financial Ratios and Prior Probabilities in Insolvency Prediction, *The Journal of Risk and Insurance*, 55, 229-244.

**Anderson, D. R. y Formisano, R. A.** (1988). Causal Factors in P-L Insolvency, *Journal of Insurance Regulation*, June, 449-461.

**Arques, A.** (1997). *La predicción del fracaso empresarial. Aplicación al riesgo crediticio bancario*, Tesis Doctoral, Universidad de Málaga.



**Bannister, J.** (1997). *Insurance Solvency Analysis*, LLP limited, Segunda Edición.

**Bar-Niv, R. y Smith, M.L.** (1987). Underwriting, Investment and Solvency, *Journal of Insurance Regulation*, 5, 409-428.

**Bar-Niv, R.** (1990). Accounting Procedures, Market Data, Cash-Flow Figures, and Insolvency Classification: The case of the Insurance Industry, *The Accounting Review*, Vol. 65, 3, July, 578-604.

**Beaver, W.H.** (1966). Financial ratios as predictors of failure. Empirical Research in Accounting: Selected Studies, *Journal of Accounting Research*, Supplement to vol. 4, 71-111.

**Best's, A.M. Company** (1991). *Best's Insolvency Study, Property and Casualty Insurers 1969-1990*.

**Browne, M. J. y Hoyt, R. E.** (1996). Economic and market predictors of insolvencies in the property-liability insurance industry, *The journal of Risk and Insurance*, 309-327.

**Bunge, M.** (1976). *La investigación científica*, Editorial Ariel.

**Calvo-Flores, A. y García, D. (coord.)** (1998a). *El riesgo financiero*, Asociación Española de Contabilidad y Administración de Empresas.

**Calvo-Flores, A. y García, D. (coord.)** (1998b). *Predicción de la Insolvencia Empresarial*, Asociación Española de Contabilidad y Administración de Empresas

**Cherkassky, V. y Mulier, F.** (1998). *Learning from data. Concepts, theory and Methods*, Wiley-Interscience Publication.

**Courtis, J.K.** (1978). Modeling a financial ratios categoric framework, *Journal of Business Finance and Accounting* 5 (4), 371-386.

**Diday E.** (ed.) (1994). *New Approaches in Classification and Data Analysis*, Springer-Verlag, Berlin.

**Dimitras, A.I.** (1995). *Multicriteria methods for the assessment of bankruptcy risk*, kPhD. Dissertation, Technical University of Crete.

**Dimitras, A.I., Zopounidis, C. y Hurson, Ch.** (1995). A multicriteria decision aid method for the assessment of business failure risk, *Foundations of Computing and Decision Sciences* 20, 2, 99-112.

**Dimitras, A.I., Zanakis, S.H. y Zopounidis, C.** (1996). A survey of business failures with and emphasis on prediction methods and industrial applications, *European Journal of Operational Research* 90, 487-513.

**Dimitras, A.I., Slowinski, R., Susmaga, R. y Zopounidis, C.** (1999). Business failure prediction using Rough Sets, *European Journal of Operational Research*, 114, 263-280.

**Directiva 91/674/CEE** del Consejo, de 19 de diciembre de 1991, *relativa a las cuentas consolidadas de las empresas de seguros*. Diario Oficial de las Comunidades Europeas. Serie Legislación 374, de 31 de diciembre.

**Dubois, D. y Prade, H.** (1992). Putting rough sets and fuzzy sets together, in *Intelligent Decision Support: Handbook of Applications and Advances in Rough Set Theory*, Slowinski, R. (ed.), Kluwer: Dordrecht 203-232.

**Eco, U.** (1983). *Cómo se hace una tesis*, Editorial, Gedisa.

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

**Fayyad, K.B. y Irani, K.B.** (1992). On the Handling of Continuous-Valued Attributes in Decision Tree Generation, *Machine Learning*, vol. 8, 87-102.

**Fernández-Palacios, J. y Maestro J.L.** (1991). *Manual de Contabilidad y Análisis Financiero*, Centro de Estudios del Seguro, S.A.

**Fernández, A.J.** *Apuntes de Gestión de Empresas Financieras*, Universidad Complutense de Madrid. Curso 1999-2000.

**Fernández, A.J.** (2000). *Análisis de los Estados Financieros de las Entidades Aseguradoras*, Aula Universitaria de Economía Actuarial y Financiera.

**Fibak, J., Pawlak, Z., Slowinski, K. y Slowinski, R.** (1986). Rough sets based decision algorithm for treatment of duodenal ulcer by HSV, *Bulletin of the Polish Academy of Sciences*, ser.Biological\_Sci. 34, nº 10-12, 227-246.

**Fibak, J., Slowinski, K. y Slowinski, R.** (1986). The application of the rough sets theory to the verification of treatment of duodenal ulcer by HSV, *Proc. 6<sup>th</sup> International Workshop on Expert Systems and their Applications*, Agence de l'Informatique, Paris, 587-599.

**Gabás, F.** (1990). *Técnicas Actuales de Análisis Contable. Evaluación de la Solvencia Empresarial*, Instituto de Contabilidad y Auditoria de Cuentas.

**Gabás, F.** (1997). Predicción de la insolvencia empresarial, en Calvo-Flores, A. y García, D. (coord.), *Predicción de la Insolvencia Empresarial*, Asociación Española de Contabilidad y Administración de Empresas, 11-32.

**García, V. y Ballester, M.** (1994). Limitaciones de los ratios financieros convencionales: una alternativa, *Actualidad Financiera*, 8/ 21-27 febrero, 159-167.

**García, V. y García-Pérez, A.** (coords.) (2000). *Decisiones Financieras y Fracaso Empresarial*, Asociación Española de Contabilidad y Administración de Empresas.

**García, D., Calvo-Flores, A. y Arques, A.** (1997) Factores discriminantes del riesgo financiero en la industria manufacturera española, en Calvo-Flores, A. y García, D. (coord.), *Predicción de la Insolvencia Empresarial*, Asociación Española de Contabilidad y Administración de Empresas, 125-156.

**Genís, R. y Almiñana, E.** (1994). Auditoria y gestión continuada: algunos aspectos relevantes, *Revista Técnica Contable*, 46 (543), 179-192.

**Greco, S., Matarazzo, B., y Slowinski, R.** (1996). Rough Approximation of Preference Relation by Dominance Relations, *ICS Research Report 16/96*, Warsaw University of Technology, Warsaw.

**Greco, S., Matarazzo, B., y Slowinski, R.** (1998). A new rough set approach to evaluation of bankruptcy risk, in C. Zopounidis (ed.), *New Operational Tools in the Management of Financial Risks*, Kluwer Academic Publishers, Dordrecht, 121-136.

**Grzymala-Busse, J.W.** (1992). LERS - a system for learning from examples based on rough sets, in: Slowinski, ed. *Intelligent Decision Support. Handbook of applications and Advances of Rough Sets Theory*, Kluwer Academic Publishers, Dordrecht, 3-18.

**Guitouni, A. y Martel, J.M.** (1998). Tentative guidelines to help choosing an appropriate MCDA method, *European Journal of Operational Research*, 109, 501-521.

**Jiménez, S., García-Ayuso, M. y Sierra, G.** (2000). *Análisis Financiero*, Editorial Pirámide.

**Karels, G.V., y Prakash, A.J.** (1987). Multivariate Normality and Forecasting of Business Bankruptcy, *Journal of Business Finance and Accounting*, Winter, 573-592.

**Komorowski, J. y Ras, Z.W.**, (eds.) (1993). *Methodologies for Intelligent Systems*. Lecture Notes in Artificial Intelligence vol. 689, Springer-Verlag, Berlin.

**Krusinska, K., Slowinski, R. y Stefanowski, J.**, (1992). Discriminant versus rough sets approach to vague data analysis, *Applied Stochastic Models and Data Analysis* 8, 43-56.

**Laitinen, E. K.**, (1992). Prediction of failure of a newly founded firm, *Journal of Business Venturing*, julio, 323-340.

**Leal, A., Sánchez-Apellániz, M., Roldán, J.L., y Vázquez, A.E.** (1999). *Decisiones empresariales con criterios múltiples. Ayudas prácticas para la dirección*, Biblioteca Eudema.

**Legislación de Seguros** (1997), Biblioteca de Legislación, 3ª edición actualizada, Civitas.

**Ley 33/1984**, de 2 de agosto, sobre *Ordenación del Seguro Privado*, BOE nº 186, 4 de agosto de 1984.

**Ley 30/1995**, de 8 de noviembre, de *Ordenación y Supervisión de los seguros privados*, BOE nº 268, 9 de noviembre de 1995.

**Lin, T. y Cercone, N.** (1997). *Rough sets and data mining: analysis of imprecise data*, Kluwer Academic Publishers. Dordrecht.

**Linares, A.** (1998). *Tratado de Auditoría de las Entidades Aseguradoras*, editorial Mapfre, S.A.

**Linares, A.**, (1998). *Contabilidad de Entidades Aseguradoras*, editorial Mapfre, S.A.

**Linares, A.** (2000). *Auditoría y Control Interno de las Entidades Aseguradoras*, Editorial Fundación Mapfre Estudios.

**Lizárraga, F.** (1996). *Modelos multivariantes de Previsión del fracaso empresarial. Una aplicación a la realidad de la información contable española*, Tesis Doctoral, Universidad Pública de Navarra.

**López, D., Moreno, J. y Rodríguez, P.** (1994). Modelos de Previsión del fracaso empresarial: aplicación a entidades de seguros, *Revista española de Seguros* 54, abril-junio, 71-110.

**Maestro, J.L.** (1998). El nuevo Plan Contable de contabilidad de Seguros, *Partida Doble*, enero.

**Maestro, J.L.** (2000). *Garantías Técnico-Financieras de las Entidades Aseguradoras*, Editorial Grupo Winterthur.

**Maroto, J.A.** (1995). Posibilidades y limitaciones del análisis económico financiero de las empresas españolas, *Papeles de Economía*, 60, 113-136.

**Martín, J.L.** (1997). Modelos de pronóstico de la insolvencia empresarial, en Calvo-Flores Segura, A. y García Pérez de Lema, D. (coord.), *Predicción de la Insolvencia Empresarial*, Asociación Española de Contabilidad y Administración de Empresas, 33-51.

**Martín, M.L., Leguey, S., Sánchez, J. M.** (1999). Solvencia y estabilidad financiera en la empresa de seguros: Metodología y evaluación empírica mediante análisis multivariante, *Cuadernos de la Fundación Mapfre Estudios*, 49.

**McKee, T.E.** (1995). Predicting bankruptcy via induction: an extension, In *Artificial Intelligence in Accounting, Finance, and Tax*, Sierra GJ, Bonson, E (eds.). J. Carrasco: Huelva, Spain.

**McKee, T.E y Lensberg, T.** (1999). Using a genetic algorithm to obtain a causally ordered model from a rough sets derived bankruptcy prediction model. Paper presented at *The International Symposium on Audit Research*, The University of Southern California.

**McKee, T.** (2000). Developing a Bankruptcy Prediction Model via Rough Sets Theory, *International Journal of Intelligent Systems in Accounting, Finance and Management*, 9, 159-173.

**Merlo, J.** y Colaborador: Revilla, M. (1983). *Contabilidad y análisis de estados financieros de entidades aseguradoras*, Editorial Mapfre, S.A.

**Michalski, R.S.** (1983). A theory and methodology of inductive learning, in *Machine Learning* (R.S. Michalski, J.G. Carbonell, T. M. Mitchell, eds), Morgan Kaufmann, 83-134.

**Mienko, R., Slowinski, R. y Stefanowski, J.** (1995). Rule Classifier Based on Valued Closeness Relation: ROUGHCLASS version 2.0, *ICS Research Report RA-95/002*, Poznan University of Technology.

**Mienko, R., Slowinski, R., Stefanowski, J. y Susmaga, R.** (1996a). RoughFamily – Software implementation of the rough set based data analysis and rule discovery techniques, in Tsumoto S. (ed), *Proceedings of the Fourth International Workshop on Rough Sets, Fuzzy Sets and Machine Discovery*, Tokyo Nov. 6-8, 437-440.

**Mienko, R., Stefanowski, J., Toumi, K y D. Vanderpooten** (1996b). Discovery-Oriented Induction of Decision Rules, *Cahier du Lamsade 141*, Paris, Université de Paris Dauphine, september.

**Millán, A.**, (2000). *Análisis Contable de Entidades Aseguradoras*, Asociación Española de Contabilidad y Administración de Empresas.

**Mora, A.** (1994). Los modelos de predicción del fracaso empresarial: una aplicación empírica del logit, *Revista Española de Financiación y Contabilidad*, 78, enero-marzo, 203-233.

**Moreno, J.** (1992). *Aplicación empírica de modelos de previsión del fracaso empresarial a entidades de seguros*, Tesis Doctoral Universidad de Sevilla.

**Moreno, A.** (2000). *La toma de decisiones en la empresa aseguradora. Un sistema dinámico aplicado a los seguros de no vida*, Tesis Doctoral, Universidad Complutense de Madrid.

**Moscarola, J.** (1978). Multicriteria decision aid – two applications in education management. In: Multiple Criteria Problem Solving (S. Zionts, ed.). *Lecture Notes in Economics and Mathematical Systems*, vol. 155, Springer-Verlag, Berlin, 402-423.

**Nurmi, H., Kacprzyk, J. y Fedrizzi, M.** (1996). Probabilistic, fuzzy and rough concepts in social choice. *European journal of Operational Research* 95, 264-277.

**Ohlson, J.A.**, (1980). Financial Ratios and the Probabilistic Prediction of Bankruptcy, *Journal of Accounting Research*, 18, 109-131.

**Orden Ministerial de 30 de julio de 1981**, por la que se aprueban las normas de adaptación del Plan General de Contabilidad a las empresas de seguros, reaseguros y capitalización.

**Orden Ministerial de 25 de marzo de 1988**, por la que se aprueba el *Real decreto 2020/1986, de 22 de agosto*, que aprueba el Reglamento de la Comisión Liquidadora de Entidades Aseguradoras, BOE nº 83, de 6 de abril.

**Pawlak, Z.** (1982). Rough Sets. *International Journal of Information & Computer Sciences* 11, 341-356.

**Pawlak, Z.** (1984). Rough classification. *International Journal of Man-Machine Studies* 20, 469-483.

**Pawlak, Z.** (1991). *Rough Sets. Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Dordrecht/ Boston/ London.



**Pawlak, Z. y Slowinski, R.** (1994). Rough set approach to multi-attribute decision analysis, *European Journal of Operational Research* 72, 443-459.

**Pawlak, Z., Slowinski, R. y Stefanowski, J.,** (1986). Rough classification of patients after highly selective vagotomy for duodenal ulcer, *International Journal of Man-Machine Studies* 24, 413-433.

**Predki, B., Slowinski, R., Stefanowski, J., Susmaga, R. y Wilk, S.** (1998). ROSE – Software Implementation of the Rough Set Theory, in L. Polkowski, A. Skowron, eds. *Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence*, vol. 1424. Springer-Verlag, Berlin, 605-608.

**Predki, B. y Wilk, S.** (1999). Rough Set Based Data Exploration Using ROSE System. In: Z.W. Ras, A. Skowron eds. *Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence*, vol. 1609, Springer-Verlag, Berlin, 172-180.

**Prieto, E.** (1973). *El Reaseguro: Función Económica*, Ediciones ICE.

**Prieto, E.** (1992). La estructura del Balance de las entidades aseguradoras en la CEE, *Revista Española de Seguros* 69, enero-marzo, 15-28.

**Prieto, E.,** (1993). Los riesgos de inversión y gestión de carteras en las entidades aseguradoras, *Análisis Financiero*, 59, 1<sup>er</sup> cuatrimestre, 95.

**Polkowski, L., Tsumoto, S. y Lin, T.** (eds.) (2000). *Rough Set methods and applications: new developments in knowledge discovery in information systems*, Physica-Verlag.

**Pozo, E.M.** (1997). *Modelos de Control de la Solvencia en Seguros no vida*. Tesis Doctoral Universidad Complutense de Madrid.

**Pozo, E.M.** (1998). Modelos de Ratios para la predicción de Insolvencias en seguros no vida, *Previsión y Seguro*, 68, 37-48.

**Real Decreto 1348/1985**, de 1 de agosto, por el que se aprueba el *Reglamento de Ordenación del Seguro Privado*. BOE nº. 185, 186, 187 de 3, 5 y 6 de agosto de 1985.

**Real Decreto 2020/1986**, de 22 de agosto, por el que se aprueba el *Reglamento de la Comisión Liquidadora de Entidades Aseguradoras*. BOE nº 235, de 1 de octubre.

**Real Decreto 2014/1997**, de 26 de diciembre, por el que se aprueba del Plan de Contabilidad de Entidades Aseguradoras y normas de formulación de las cuentas de los grupos aseguradores, BOE, 30 de diciembre de 1997.

**Real Decreto 2486/1998**, de 20 de noviembre, por el que se aprueba el *Reglamento de Ordenación y supervisión de los Seguros Privados*. BOE nº 282 de 25 de noviembre de 1998.

**Rivero, P.** (1987). *Análisis de balances y estados complementarios*, Editorial Pirámide.

**Rivero, P., Banegas, R., Sánchez, F. y Nevado, D.** (1998). *Análisis por Ratios de los Estados Contables Financieros (Análisis Externo)*, Editorial Civitas.

**Rodríguez, M.C.** (1990). *La Predicción de las Crisis Empresariales. Modelos para el sector Asegurador*, Secretariado de Publicaciones Universidad de Valladolid.

**Rodríguez, F. y Bravo, F.** (1998). La auditoria de las Entidades Aseguradoras, *Partida Doble* 94, noviembre, 38-43.

**Rodríguez-Vilariño, M.L.** (1994). Utilidad del Análisis de Ratios para la predicción de la Insolvencia Empresarial, (I), (II) y (III), *Actualidad Financiera*, 34, 699-724; 35, 725-750; 36, 751-773.

**Romanski, S.** (1988). Operation on families of sets for exhaustive search, given a monotonic function, in W. Beeri, C.Schmidt, N. Doyle (eds.), *Proceedings of the third Int. Conference on Data and Knowledge Bases*, Jerusalem, 310-322.

**Romero, C.** (1993). *Teoría de la decisión multicriterio: Conceptos, técnicas y aplicaciones*, Alianza Universidad Textos.

**Roqueta, A.** (2000). El principio de empresa en funcionamiento: a través de un ratio financiero, *El Auditor* 8- Diciembre 2000, 2-5.

**Roubens, M. y Vincke, Ph.** (1985). Preference Modeling. *Lecture Notes in Economics and Mathematical Systems*, vol. 250, Springer-Verlag, Berlin.

**Roy, B.** (1985). *Méthodologie Multicritère d'Aide à la Décision*. Economica, Paris.

**Roy, B.** (1989). Main sources of inaccurate determination, uncertainty and imprecision in decision models. *Math. Comput. Modell.*, 12, 1245-1254.

**Roy, B.** (1992). Decision science or decision aid science, *European Journal of Operational Research*, Special Issue on Model Validation in Operations Research.

**Roy, B.** (1996). *Multicriteria Methodology for Decision Aiding*, Kluwer Academic Publishers, The Netherlands.

**Roy, B. y Bouyssou D.** (1993). *Aide Multicritère à la Décision: Méthodes et Cas*. Economica, Paris.

**Sanchis, A.** (2000). *Una aplicación del Análisis Discriminante a la previsión de la Insolvencia en las empresas españolas de seguros no-vida*, Tesis Doctoral, Universidad Complutense de Madrid.

**Sanchis, A., Gil, J.A. y Heras, A.** (2002). El análisis discriminante en la previsión de la insolvencia en las empresas de seguros no vida, *Revista Española de Financiación y Contabilidad* (to appear).

**Segovia, M.J., Gil, J.A., Heras, A., Vilar, J.L. y Sanchis, A.** (2002). Using Rough Sets to predict insolvency of Spanish non-life insurance companies, *Proceedings of 6<sup>th</sup> International Congress on Insurance: Mathematics and Economics*, Lisboa.

**Segovia, M.J., Gil, J.A., Heras, A. y Vilar, J.L.** (2002). Predicción de insolvencias con el método Rough Set, *X Jornadas de Asepuma*, Madrid.

**Serrano, C. y Martín del Brío, B.**, (1993). Predicción de la Crisis Bancaria Mediante el Empleo de Redes Neuronales Artificiales, *Revista Española de Financiación y Contabilidad XXII*, nº 74, p.153-176.

**Shaked, I.** (1985). Measuring Prospective Probabilities of Insolvency: An Application to the life Insurance Industry, *The Journal of Risk and Insurance* 52, 59-80.

**Siegel, Ph., de Korvin, A. y Omer, K.** (1993). Detection of irregularities by auditors: A rough set approach, *Indian Journal of Accounting*, 44-54.

**Sierra, R.** (1988). *Tesis doctorales y trabajos de investigación científica*, Editorial Paraninfo, S.A.

**Skowron, A.** (1991a). *The Implementation of Algorithms based on Discernibility Matrix*, Manuscript.

**Skowron, A.** (1991b). The Discernibility Matrices and Functions in Information Systems, *Institute of Computer Sciences Reports*, Warsaw Technical University and Fundamental Informaticae.

**Skowron, A.** (1993). Boolean reasoning for decision rules generation, in: J. Komorowski and Z.W. Ras (eds.), *Methodologies for Intelligent Systems*, Lecture Notes in Artificial Intelligence vol. 689, Springer-Verlag, Berlin, 295-305.

**Skowron A. y Rauszer C.** (1992). The discernibility matrices and functions in R. Slowinski (ed.), *Decision Support by Experience – Applicants of The Rough Sets Theory*. Kluwer Academic Publishers, Dordrecht, 331-362.

**Slowinski, R.** ed. (1992). *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, Kluwer Academic Publishers, Dordrecht-The Netherlands.

**Slowinski, R.** (1993). Rough set learning of preferential attitude in multicriteria decision making, in: J. Komorowski and Z. W. Ras (eds.), *Methodologies for Intelligent Systems*. Lecture Notes in Artificial Intelligence vol. 689, Springer-Verlag, Berlin, 642-651.

**Slowinski, R.** (1995). Rough set approach to decision analysis, *AI Expert Magazine* 10, 3, 18-25.

**Slowinski, K. y Slowinski, R.** (1989). Sensitivity analysis of rough classification, *International Journal of Man-Machine Studies*, 30.

**Slowinski, K., Slowinski, R. y Stefanowski, J.** (1988). Rough Sets Approach to analysis of data from peritoneal lavage in acute pancreatitis, *Medical Informatics* 13, 3, 143-159.

**Slowinski, R. y Stefanowski, J.** (1989). Rough classification in incomplete information systems, *Mathematical and Computer Modeling*, 12, 1347-1357.

**Slowinski R. y Stefanowski J.** (1992). "RoughDAS" and "RoughClass" software implementations of the rough set approach. Chapter III-8 in Slowinski (1992), 445-456.

**Slowinski R. y Stefanowski J.** (1994a). *RoughDas: Rough Set Based Data Analysis System-Version 2.0- User's Guide book*, Poznan, Poland.

**Slowinski R. y Stefanowski J.** (1994b). Rough classification with valued closeness relation, in: Diday E. et al (eds.), *New Approaches in Classification and Data Analysis*, Springer-Verlag, Berlin, 482-488.

**Slowinski R. y Vanderpooten, D.** (1997). Similarity relation as a basis for rough approximations, in P.P. Wang (ed.): *Advances in Machine Intelligence & Soft-Computing*, Raleigh, NC, 17-33.

**Slowinski R. y Zopounidis, C.** (1995). Application of the rough set approach to evaluation of bankruptcy risk, *International Journal of Intelligent Systems in Accounting, Finance and Management* 4, nº 1, 27-41.

**Slowinski R., Zopounidis, C. y Dimitras, A.I.** (1997). Prediction of company acquisition in Greece by means of the rough set approach, *European journal of Operational Research* 100, nº1, 1-15.

**Stefanowski, J.** (1992). Rough Sets Theory and discriminant methods as tools for analysis of information systems. A comparative study, *Foundations of Computing and Decision Sciences* vol 17 (2), 81-98.

**Stefanowski J.** (1993a). Classification and decision supporting based on the rough set theory, *Foundations of Computing and Decision Sciences*, vol. 18, nº 3-4, 371-380.

**Stefanowski J.** (1993 b). Classification support based on the rough set theory, in J. Wessels, A.P. Wierzbicki (eds.): *User-Oriented*

---

*Methodology and Techniques of Decision Analysis and Support, LNEMS 397*, Springer Verlag, Berlin, 185-192.

**Stefanowski, J. y Vanderpooten, D.** (1994). A general two-stage approach to inducing rules from examples. In: Ziarko, W. (Ed.), *Rough Sets, Fuzzy Sets and Knowledge Discovery*, Springer, London, 317-325.

**Tam, K.Y. y Kiang, M.Y.** (1992). Managerial applications of neural networks: The case of bank failure predictions, *Management Science*, vol. 38, nº 7, 926-947.

**Trieschmann, J. y Pinches, G.** (1973). A multivariate Model for Predicting Financially Distressed P-L Insurers, *Journal of Risk Insurance*, nº 40, 327-338.

**Urias, J.** (1999). *Análisis de Estados Financieros* (Segunda Edición), Editorial McGraw-Hill.

**Uriel, E.** (1995). *Análisis de Datos. Series Temporales y Análisis Multivariante*, Colección Plan Nuevo AC.

**Vilar, J.L.** (1990). *Fundamentos matemáticos del estudio de la solvencia dinámica en seguros no vida*, Tesis Doctoral, Universidad Complutense de Madrid.

**Zavgren, C.V.** (1985). Assessing the vulnerability to failure of American industrial firms. A logistic analysis, *Journal of Business Finance and Accounting* 12 (1), 19-45.

**Ziarko, W.** (1993). Analysis of Uncertain Information in The Framework of Variable Precision Rough Sets. *Foundations of Computing And Decision Sciences* Vol. 18, Nº 3-4, 381-396.

**Ziarko, W. (ed.)** (1994). *Rough sets, Fuzzy Sets and Knowledge Discovery*, Springer-Verlag, Berlin.

**Ziarko, W., Golan, D. y Edwards, D.** (1993). An application of DATALOGIC/R knowledge discovery tool to identify strong predictive rules in stock market data, in *Proceedings of AAAI Workshop on Knowledge Discovery in Databases*, Washington D.C., 89-101.

**Zopounidis, C.** (1995). Evaluation du risque de défaillance de l'entreprise: Méthodes et cas d'application *Economica*, Paris.

**Zopounidis, C.** (1999). Multicriteria decision aid in financial management, *European Journal of Operational Research*, 119, 404-415.

**Zopounidis, C. y Dimitras A.** (1998). *Multicriteria Decision Aid Methods for the Prediction of Business Failure*, Kluwer Academic Publishers, Dordrecht.