UNIVERSIDAD DE GRANADA

E.T.S. DE INGENIERÍAS INFORMÁTICA Y DE TELECOMUNICACIÓN

Departamento de Ciencias de la Computación e Inteligencia Artificial

UNA NUEVA PERSPECTIVA PARA RECUPERACIÓN EN RAZONAMIENTO BASADO EN CASOS: MEJORA DE LA ADECUACIÓN DEL CASO RECUPERADO USANDO FUNCIONES DE RIESGO

Tesis Doctoral

María Isabel Navarro Jiménez

Directores: J.L. Castro Peña, J.M. Zurita López

Granada, Octubre de 2011

Editor: Editorial de la Universidad de Granada
Autor: María Isabel Navarro Jiménez
D.L.: GR 1179-2012
ISBN: 978-84-695-1188-6

La memoria titulada "Una Nueva Perspectiva para Recuperación en Razonamiento Basado en Casos: Mejora de la Adecuación del Caso Recuperado Usando Funciones de Riesgo", que presenta D^{a}. María Isabel Navarro Jiménez, para optar al grado de Doctor, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección de los Doctores Juan Luis Castro Peña y José Manuel Zurita López.

Granada, Octubre de 2011

El doctorando

M.I. Navarro Jiménez

Los directores

J.L. Castro Peña

J.M. Zurita López

Agradecimientos

Para comenzar, quiero dirigirme a mi familia y amigos para agradecerles su apoyo, paciencia y comprensión en este largo periodo de ausencia y mal humor. Siento haber sido insoportable y no haber participado de momentos importantes. Me gustaría dedicarlo, especialmente a mi abuela, a la que estoy segura que le hubiera gustado mucho poder ver el final, y en particular agradecer a mi padre su esfuerzo.

También decir a todos mis compañeros, sobre todo a los que son miembros del grupo de investigación, gracias por su incondicional ayuda.

Agradecer, como no, a mis directores de Tesis, Juan Luis y José Manuel, la paciencia que han tenido conmigo y el apoyo que me han ofrecido, tanto en el plano profesional como en el personal. Si alguien ha sido decisivo en la realización de este trabajo, sin duda, han sido ellos.

Finalmente comentar que son tantas las personas que se han visto implicadas en este largo proceso de una forma u otra y son tantas a las que me gustaría dar mi agradecimiento, que resulta casi imposible nombrarlas. A todos ellos, GRACIAS.

Índice general

Índice de cuadros

Índice de figuras

Capítulo 1

Introducción

El objetivo de este capítulo es presentar esta memoria. Para ello, veremos: la motivación, una descripción de los objetivos que pretendemos conseguir y la forma en que está estructurada.

1.1. Motivación

El Razonamiento Basado en Casos (RBC) es una técnica de resolución de problemas muy conocida dentro de la Inteligencia Artificial (IA), que resuelve problemas imitando el modo en que razonan los humanos. Su uso actualmente está muy extendido, y aunque está siendo utilizado con éxito para resolver numerosos problemas en muy diversos contextos, encontramos aún ciertos aspectos en los actuales modelos de RBC que aún pueden ser mejorados para avanzar en esa conducta inteligente. Cuando un sistema de RBC resuelve un problema, generalmente lo que hace es buscar el más similar de entre todos los que tiene en memoria, y finalmente decide aplicar la solución de ese problema a la resolución del nuevo. Sin embargo, cuando nosotros resolvemos un problema, (no sólo buscamos el más similar y aplicamos la solución de este problema para resolverlo), sino que antes de hacerlo consideramos las consecuencias futuras que conlleva aplicar dicha solución. Estas consecuencias son fundamentales, ya que aportan una información muy importante y en ocasiones determinante a la hora de decidir. El estudio e incorporación en un modelo de RBC de esta información es la motivación principal de esta investigación.

Siguiendo este criterio se puede afirmar que en la vida real nos enfrentamos a dos tipos diferentes de situaciones o problemas. Aquellas en las que resultarán

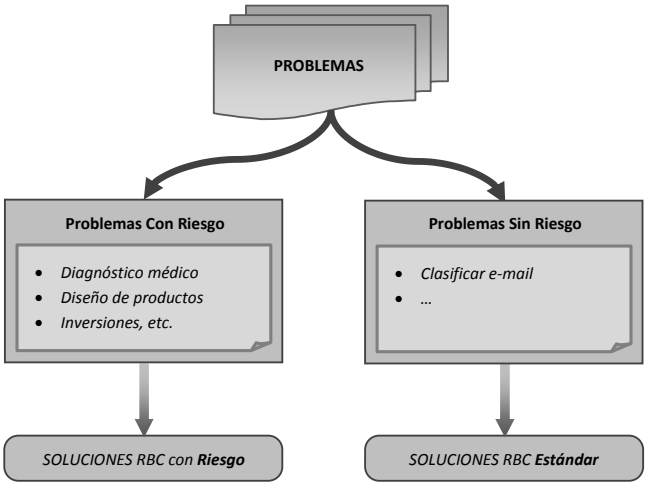


Figura 1.1: Tipos de problemas

determinantes las consecuencias de una mala decisión y aquellas en las que éstas consecuencias no lo sean. Por ejemplo, al diagnosticar una enfermedad, sería conveniente tener en cuenta, las consecuencias que conllevaría un posible diagnostico equivocado, puesto que la vida del paciente podría estar en juego. Este sería un claro ejemplo, de situación o problema, en la que una mala decisión es determinante. También existirán otras situaciones, en las que no sea tan relevante esta información, por ejemplo, clasificar correo spam, que aunque conlleva cierto riesgo asociado, las consecuencias de una mala clasificación no son tan cruciales. Luego, se puede afirmar que existen dos clases de problemas claramente diferenciados (ver Figura 1.1):

a) *Problemas Con Riesgo*(PCR), son problemas donde tomar una decisión conlleva un riesgo asociado significativo.

b) *Problemas Sin Riesgo*(PSR), problemas en los que el riesgo es muy pequeño, poco significativo o incluso inexistente.

Para resolver problemas que pertenecen a la clase *Problemas Sin Riesgo*, se han desarrollado numerosas técnicas. Todas ellas tienen en común que trabajan de forma similar: primero, calculan la similitud entre los atributos del problema a resolver, con respecto a los atributos de los problemas almacenados en memoria; luego, se calcula la media global ponderada por la importancia que cada atributo tiene dentro del problema. Aunque estas técnicas han resuelto con éxito multitud de problemas, son susceptibles de mejora para para enfrentarse a situaciones que conllevan asociado

un riesgo alto, por ejemplo: problemas de dinero o salud. Al tratar problemas de salud, como se ha comentado, es muy importante considerar si la solución es peligrosa para el paciente, puesto que no se debe poner en riesgo su vida. Igualmente en problemas financieros, donde está en juego dinero, se deben tener en cuenta las consecuencias de la decisión tomada, puesto que una mala decisión podría provocar un gran quebranto económico.

Por tanto, parece razonable, que a la hora de resolver un problema que pertenezca a la clase *Problemas Con Riesgo*, se tengan en cuenta las consecuencias de aplicar una solución no adecuada y así poder evitar cometer un error con consecuencias fatales. Esto hace que sea interesante tomar esta información como parte del problema. Llegados a este punto, cabe plantearse los siguientes interrogantes:

- ¿Qué criterios serán los adecuados para elegir el caso más conveniente y no solo el más similar?
- ¿Cómo se puede introducir esta información dentro del problema?
- ¿Cómo se comportará esta metodología comparada con otras ya existentes?
- ¿Qué ventajas aporta? ¿Cuáles serán sus debilidades?
- ¿Puede aplicarse a problemas reales?

1.2. Objetivos

Esta memoria se centra en el estudio de la etapa de Recuperación en Razonamiento Basado en Casos. Persigue como objetivo general aumentar la calidad de los sistemas RBC, evitando cometer errores que conlleven consecuencias no deseadas en esta etapa. Para ello, se define una nueva metodología que está orientada específicamente a problemas de la clase PCR. Este objetivo general se concreta en los siguientes objetivos específicos, como muestra la Figura 1.2:

- *Analizar los métodos de recuperación existentes*: En este punto se plantea el estudio de los métodos de recuperación más empleados en la literatura, o cuyas técnicas resulten más novedosas con el fin de realizar una propuesta adecuada para problemas con riesgo; y de forma, que pueda apreciarse su originalidad.

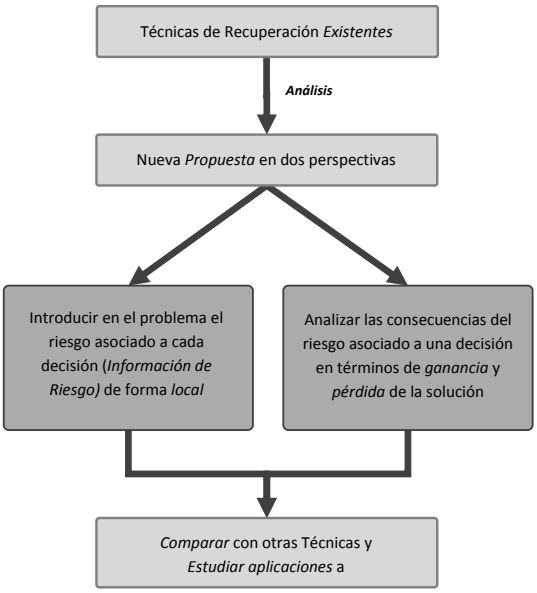


Figura 1.2: Objetivos

- *Mejorar la bondad de la recuperación de un caso, introduciendo en el problema la cantidad de riesgo asociada a cada decisión*: Con este objetivo nos proponemos desarrollar una nueva metodología orientada a resolver problemas que pertenezcan a la clase PCR. Aplicando esta metodología, ahora, no solo se recuperará el caso más similar, sino que se recupera el caso cuya solución sea la más adecuada para resolver el problema. Esta nueva información se llamará *Información de Riesgo* y será la encargada de indicar o medir lo adecuada que resulta la solución asociada al caso considerado para resolver el problema. Para controlar que todos los valores son fiables se utilizará el conocimiento de un experto. Esta nueva información se introduce en el proceso de resolución del problema añadiendo un nuevo paso o etapa. A esta nueva etapa (o paso) se la llamará *Adecuación* de la medida.

- *Proponer modelos para definir la función Información de Riesgo en ausencia total o parcial del experto*: Con este objetivo se persigue desvincularse del experto, puesto que el depender constantemente de él es una situación costosa y en algunas ocasiones poco realista. Para ello, se proponen y analizan distintos modelos de asignación de riesgo. Algunos de estos modelos son ya conocidos y serán adaptados al problema. También se define uno propio para cubrir el caso en que ninguno de estos funcione o puedan aplicarse. Los resultados obtenidos

por los modelos propuestos, serán comparados entre sí y con los resultados obtenidos cuando la información proviene en su totalidad del experto, para así estudiar su comportamiento y obtener conclusiones y criterios sobre cuando será adecuado usar cada uno de ellos.

- *Analizar la influencia de las consecuencias asociadas a cada decisión en términos de ganancia y pérdida.* En este punto, se analizan más en profundidad las consecuencias de seleccionar una u otra solución. Como ya se ha comentado, elegir una solución, sin tener en cuenta sus consecuencias, no es del todo congruente con el proceso de razonamiento humano. Resulta más consecuente con este proceso, que antes de decidir qué caso recuperar, su solución asociada sea ponderada en términos de su ganancia y pérdida potencial. Sobre todo teniendo en cuenta que en la vida real la *ganancia* y la *pérdida* no son conceptos complementarios, es decir, *ganancia* $\neq$ - *pérdida*. Luego, con este objetivo, se pretende desarrollar una metodología cuya ventaja reside en minimizar el coste de cada decisión tomada por el sistema. Para ello, se definen los conceptos de *ganancia* y *pérdida* de una solución, como dos funciones que medirán las consecuencias positivas y negativas de cada decisión. Más tarde, en términos de la ganancia y la pérdida se obtendrá el *beneficio* de dicha decisión. Finalmente, la información resultante se introduce en el proceso de selección del caso a través de un sistema de inferencia difuso.

- *Comparar empíricamente los modelos propuestos con otros ya existentes en la literatura.* Realizar pruebas, para obtener resultados experimentales y así comparar la metodología propuesta en este trabajo con otros métodos propuestos en la literatura para ver su funcionamiento y obtener conclusiones.

- *Desarrollar aplicaciones a problemas reales de la nueva metodología presentada.* Finalmente, y para dar mayor validez a la investigación realizada, se adapta la metodología presentada a dos problemas reales.

1.3. Estructura del trabajo

La presente memoria consta de siete capítulos y tres apéndices. Se procede a continuación a la descripción resumida de cada uno de ellos.

Capítulo 1. *Introducción.* Es el capítulo en que se plantea el trabajo, se muestran los objetivos a conseguir y un breve resumen de cómo se ha organizado la memoria.

Capítulo 2. *RBC y la Etapa de Recuperación.* A lo largo de este capítulo se describe el marco teórico en el que se desarrolla esta propuesta. Para ello, se estudia el funcionamiento general de un razonador basado en casos, su historia y las ventajas e inconvenientes de su uso. También se analiza en detalle la etapa de recuperación y el concepto de similitud. Finalmente, se presenta una visión en conjunto de las investigaciones más relevantes que tratan la recuperación como tema principal.

Capítulo 3. *Un Nuevo Concepto en Recuperación: la Información de Riesgo.* En este capítulo se define un nuevo concepto que sirve como refuerzo en la recuperación y al que llamamos *Información de Riesgo*. Este concepto, ayudará al sistema a recuperar no sólo el caso más similar al problema, si no también el caso más adecuado para resolverlo. Se definido de forma local, atributo por atributo, y sus valores son asignados con ayuda del experto. Finalmente, se aplica el modelo a diferentes bases de datos y se comparan los resultados con los obtenidos utilizando conocidas técnicas de recuperación.

Capítulo 4. *Modelos para Definir la Función Información de Riesgo.* Una vez se han visto las mejoras que aporta el considerar la información riesgo para resolver un problema, se propone como objetivo hacer más eficiente su uso. Para ello se evitará todo lo posible la dependencia del experto, minimizando así los costes del sistema. Se proponen métodos que aproximen la función Información de Riesgo en distintos escenarios y se realiza un estudio de cuando será más adecuado utilizar cada método. Se persigue que el modelo presentado en el capítulo 3, funcione tanto en ausencia parcial como en ausencia total del experto; consiguiendo que aprenda la información que éste aportó anteriormente. También se desarrollará un método propio capaz de asignar el riesgo cuando no se tenga información alguna del experto, este método al no contar con información del exterior, asignará el riesgo en función de la información que le aporten los casos almacenados en la base de casos.

Capítulo 5. *Funciones de Ganancia y Pérdida para Recuperación en RBC.* En este capítulo se pretende dar otro paso y así profundizar aún más en el concepto de riesgo. Teniendo en cuenta que el riesgo puede provenir de distintas características del problema, dividiremos dicho concepto en *Ganancia* y *Pérdida* de cada solución. Por eso, antes de recuperar un caso, ya no sólo se tendrá en cuenta el riesgo de que la solución sea o no conveniente, sino la pérdida o ganancia que cada solución aporta, es decir, ahora la decisión se comparte. Esto es con-

veniente, puesto que en la mayoría de los problemas reales, la ganancia no es complementaria a la pérdida (es decir, *ganancia* $\neq 1-$*pérdida*). Finalmente el modelo se aplica a diferentes bases de datos y los resultados se comparan con los obtenidos utilizando conocidas técnicas de recuperación.

Capítulo 6. *Aplicaciones a Problemas Reales.* Este capítulo muestra dos aplicaciones a problemas reales de la metodología presentada en esta memoria. En la primera aplicación utilizando los conceptos de Información de Riesgo y de probabilidad de una solución definidos en los capítulos 3 y 5 respectivamente, se desarrolla un sistema que a partir de datos reales sobre la flexibilidad y estrategia de operaciones en 72 empresas de consultoría españolas, obtiene conclusiones en forma de reglas lingüísticas, sobre cómo éstas se relacionan. La otra aplicación adapta el concepto de Información de Riesgo al desarrollo de nuevos productos. La aplicación consiste en un sistema de RBC que ayuda a diseñar nuevos productos. Este sistema se construyó junto con un experto de la empresa Beltone, para el diseño de un nuevo audífono.

Capítulo 7. *Resumen, Conclusiones y Trabajos Futuros.* Se realiza un resumen de los aspectos más relevantes del trabajo, se presentan las conclusiones finales del mismo y se analizan las líneas de trabajo más prometedoras para su continuación, con el objetivo de mejorar o extender de algún modo las propuestas introducidas en los capítulos anteriores.

Por último se incluyen tres apéndices: uno, en el que a modo de ejemplo, se ilustra como trabaja la metodología presentada en el capítulo 3, otro, en el que puede verse gráficamente cómo se comportan los modelos presentados en el capítulo 4 y un apéndice final, con el cuestionario empleado para recoger los datos utilizados en una de las aplicaciones del capítulo 5.

Capítulo 2

El RBC y la Etapa de Recuperación

Razonamiento Basado en Casos (RBC), es una técnica relativamente reciente de resolución de problemas, que cada vez atrae más la atención de investigadores y empresas. Son múltiples las posibilidades que aporta en diversos campos ya que es una herramienta fácil de usar y muy intuitiva, que además, permite aprendizaje. En este contexto, el principal objetivo de este capítulo es dar una descripción sobre RBC, haciendo una breve exposición de su historia y funcionamiento. Finalmente, se presentará un estado del arte de las medidas de similitud y técnicas de recuperación más relevantes.

2.1. Introducción

La experiencia en la resolución de problemas es una cualidad que poseen los humanos y que es necesario aprender para poder crear un modelo de comportamiento inteligente. Este es el principio en que se basa el Razonamiento Basado en Casos, ya que para resolver problemas nuevos, adapta soluciones de experiencias similares previas. Esto le ha convertido en una técnica popular de resolución de problemas, que además, permite aprendizaje.

El gran interés que ha suscitado desde su nacimiento lo ha convertido en tema central de muchas investigaciones y conferencias. Tres workshops sobre RBC fueron organizados en 1988, 1989 y 1991 por la Agencia americana DARPA (Defense Advanced Research Projects Agency). Fueron estos, los que precisamente marcaron el nacimiento de esta disciplina. Fue en 1993 cuando tuvo lugar el primer workshop

sobre RBC en Europa (European Workshop on Case-based Reasoning, EWCBR-93) con más de 120 participantes. Desde entonces multitud de workshops internacionales y conferencias sobre RBC han tenido lugar en diferentes partes del mundo. La primera conferencia internacional sobre RBC (International Conference on Case-based Reasoning, ICCBR-95) se celebró en Sesimbra (Portugal). Actualmente continúa celebrándose y en 2011 ha tenido lugar en Londres (Reino Unido).

Básicamente, el funcionamiento de un razonador basado en casos, tiene cuatro etapas claramente diferenciadas: la etapa de *recuperación* donde se buscan casos similares al problema, la de *adaptación* dónde se adecúa la solución recuperada para así poder resolver el problema, la de *evaluación y revisión* donde se evalúa y revisa la bondad de la solución encontrada y la etapa de *aprendizaje* en la que se aprende de la resolución tanto si ésta ha tenido éxito como si no. De entre estas etapas la *recuperación* juega un papel muy importante, puesto que si el caso recuperado no es el adecuado, la solución propuesta no servirá para resolver el problema y por tanto, habría que repetir todo el proceso. Esta investigación se centra justamente en esta etapa y su objetivo es aumentar la calidad de los sistemas RBC, evitando consecuencias erróneas o no deseadas en su funcionamiento. Por lo tanto, a lo largo de este capítulo, se dará una noción de los conocimientos básicos necesarios. Para profundizar sobre el tema se recomienda consultar la siguiente bibliografía [125, 89].

El resto del Capítulo está organizado de la siguiente forma: la Sección 2 es una introducción general al Razonamiento Basado en Casos, en la que se hará un breve recorrido sobre su historia, se estudiará de forma detallada el ciclo del RBC y sus características más relevantes. La Sección 3, presenta una introducción al concepto de medida de similitud y un repaso a las medidas más utilizadas. También se revisaran, de forma análoga, algunas de las técnicas de recuperación más significativas en RBC. Finalmente en la Sección 4, se expondrán tanto un resumen como las conclusiones del capítulo.

2.2. Razonamiento Basado en Casos

A lo largo de esta sección se estudiará cómo funciona un razonador basado en casos, cuáles son las principales etapas que lo componen y cómo ha sido la evolución de este tipo de sistemas.

2.2.1. Historia del Razonamiento Basado en Casos

Se pueden encontrar muchas referencias significativas y de muy diversos autores sobre los orígenes del RBC. Aunque en lo que parece haber consenso, es que fue el trabajo de Roger Schank y su grupo de investigación de la Universidad de Yale [147, 146], a principios de los ochenta, quienes realizaron el primer modelo cognitivo para RBC y la primera aplicación basada en este modelo. Partieron de la idea de que el conocimiento que nosotros tenemos sobre las distintas situaciones a las que nos enfrentamos cotidianamente se guarda en nuestra mente en forma de recuerdos. Y son estos recuerdos los que nos proporcionan la posibilidad de inferir conclusiones. La constante investigación de Roger Schank sobre el papel que este recuerdo de situaciones previas podía jugar en la resolución de problemas y en el aprendizaje, derivó en el primer modelo dinámico de memoria.

Sin embargo, fue Janet Kolodner quien desarrolló el primer sistema RBC implementando el modelo dinámico de memoria de Schank. El sistema llamado CYRUS [87, 88] debido a Cyrus Vance ex-secretario de los Estados Unidos, contenía en cada caso información sobre sus viajes y reuniones. Este mismo modelo de memoria también fue la base de sistemas como: CASEY [91], que genera explicaciones sobre los síntomas de enfermos de corazón; JULIA [69], que diseña menús, adaptando, transformando e imponiendo restricciones nuevas a menús previamente diseñados; PERSUADER [161], que genera soluciones para negociaciones laborales que sean aceptables para ambas partes, etc.

Más tarde un método alternativo al modelo de memoria planteado por Schank, vino con el trabajo realizado por Bruce Porter (Universidad de Texas). Quien utilizando clasificaciones heurísticas y métodos de machine learning creó el sistema PROTOS [132], que se ha aplicado para el aprendizaje de trastornos en el oído.

Tras estas exitosas aplicaciones, no resultó sorprendente el hecho de que se utilizara RBC en el mundo legal, ya que la practica de la ley esta basada en la consideración de precedentes y en la noción de caso. Fue el grupo de trabajo dirigido por Edwina Rissland de la Universidad de Massachusetts, quien desarrolló HYPO [13], sistema que genera argumentos legales citando casos pasados a favor y en contra como justificación de sus argumentos. Más tarde, este sistema fue combinado con razonamiento basado en reglas para producir CABARET [139].

Como se ha visto, las primeras investigaciones realizadas sobre RBC provienen principalmente de EEUU, fue un poco más tarde cuando tomó auge en Europa. Entre los primeros grupos de investigación europeos se encuentra el de Derek Sleeman, Aberdeen (Escocia), el cual desarrolló el sistema REFINER [149], estudiando el uso de casos para adquisición de conocimiento. Casi al mismo tiempo, y también en Europa Michael Ritcher y Klaus Althoff [8], Universidad de Kaiserslautern (Alemania), aplicaron RBC a diagnósticos complejos, creando así el sistema PATDEX [138]. Por otro lado Agnar Aamodt, Universidad de Trondheim (Noruega) investigó sobre el aprendizaje con RBC combinando los casos con el dominio general del conocimiento, resultando el sistema CREEK [1, 2] muy similar al PROTOS. Siguiendo su desarrollo en Europa, en Reino Unido se generaron aplicaciones de RBC a la ingeniería civil. Un grupo de la Universidad de Salford lo aplicó a diagnósticos fallidos, para reparación y rehabilitación de edificios [176] y Robertson y Yang (Edimburgo) [190] a la construcción. En España, cabe destacar, los siguientes grupos: GAIA - Group for Artificial Intelligence Applications (Universidad Complutense, Madrid), BISITE - Bioinformática, Sistemas Inteligentes, Tecnología Educativa (Universidad de Salamanca), eXiT - Ingeniería de Control y Sistemas Inteligentes (Universidad de Girona), Instituto de Investigación en Inteligencia Artificial - IIIA (Consejo Superior de Investigaciones Científicas, CSIC), entre otros.

Actualmente el RBC es una técnica estudiada ampliamente [96] que se ha extendido por todo el mundo y que tiene multitud de aplicaciones en muy diversos dominios que van desde control de tráfico [142], predecir problemas financieros [102], clasificación de clientes [6], hasta diagnosticar enfermedades [72], e-comercio [196], cocina [47], filtrar correo spam [49] o la agricultura [191]. Puede decirse que parte del éxito del RBC se debe principalmente al incremento del número de artículos en revistas del ámbito de la IA y también al incremento del número de aplicaciones comerciales de RBC con éxito. Todo esto ha hecho que cada vez más países e investigadores demuestren un interés activo por él.

2.2.2. Funcionamiento de un sistema RBC

Un sistema de RBC resuelve problemas haciendo uso de la experiencia. De forma, que cuando el sistema se enfrenta a un problema, recuerda soluciones que funcionaron bien con problemas similares y las utiliza como punto de partida en la resolución de dicho problema. Por ejemplo, si consideramos la forma en que un

médico efectúa un diagnóstico, estaremos ante un típico ejemplo de como trabaja un sistema de RBC. El médico tiene una serie de experiencias almacenadas, que se corresponden con las enfermedades de los distintos pacientes a los que ha tratado a lo largo de su carrera. Cuando llega un nuevo paciente a su consulta lo compara con pacientes anteriores que tuvieron síntomas similares. El tratamiento utilizado para curar a aquellos pacientes, es entonces, usado y modificado, si fuera necesario, de forma conveniente para tratar al nuevo paciente.

En la vida real hay muchas situaciones en las que se emplea RBC, por ejemplo: cuando se cocinan tallarines, ya que se hace en base a como se prepara cualquier otro tipo de pasta; cuando se intenta averiguar el precio de una vivienda, para ello se usan informaciones provenientes de otras propiedades similares, etc. Por lo que se puede afirmar que casi siempre es más fácil la segunda vez que se resuelve un problema que la primera, ya que se tiene la posibilidad de recordar lo que se hizo y repetir la solución que ya se utilizó con buenos resultados.

Formalmente, un sistema de RBC resuelve nuevos problemas adaptando soluciones de problemas antiguos. Por tanto, el RBC implica razonamiento a través de experiencias previas. Para ello, retiene en memoria los problemas tratados anteriormente junto con sus respectivas soluciones, para así, resolver nuevos problemas utilizando como referencia ese conocimiento almacenado. Generalmente, cuando un sistema RBC se encuentre con un nuevo problema, éste buscará en su memoria de problemas pasados, a la que en adelante se llamará *base de casos*, e intentará encontrar el problema, *caso*, cuyas características sean iguales o similares a las del problema tratado. Si el razonador no puede encontrar un caso idéntico en su base de casos, entonces intentará buscar uno o varios casos que sean lo más parecidos posible al caso actual. En situaciones donde el caso recuperado es idéntico al problema que nos ocupa, y asumiendo que la solución con la que se almacenó fue un éxito, ésta es propuesta como solución para resolverlo. Si lo que se recupera, no es un caso idéntico, entonces se adapta la solución. Para ello se identifican las diferencias entre ambos casos y se modifica la solución del caso recuperado tomando en cuenta dichas diferencias.

La estructura de un razonador basado en casos, a un alto nivel de abstracción, puede verse como muestra la Figura 2.1. Dicho sistema incorpora los siguiente factores externos:

- Los detalles de entrada del problema o *caso*.

- La salida del sistema o *solución*.
- La memoria de los casos pasados o *base de casos*, la cual es utilizada por el *mecanismo de razonamiento*, para poder obtener la solución correcta.

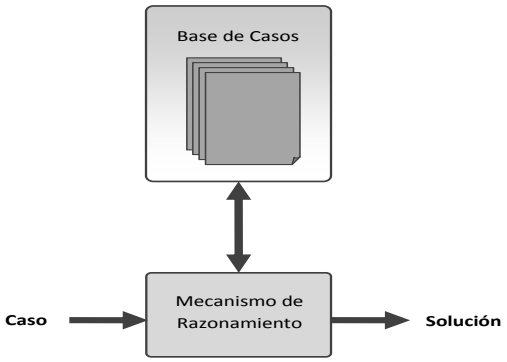


Figura 2.1: Esquema general de funcionamiento de un RBC

Normalmente la estructura interna del mecanismo de razonamiento, que en adelante se llamará *razonador*, está divida en dos partes principales que son: el *recuperador de casos* y el *razonador* de casos. Mientras que la tarea del *recuperador de casos* es la de encontrar el caso más apropiado de la base de casos, la del *razonador de casos* es usar el caso recuperado para encontrar la solución del problema. Cuando el caso recuperado es el mismo que se introdujo como entrada, no es necesario el razonamiento, puesto que el propio caso recuperado contiene directamente la solución del caso actual; aunque hay que hacer notar que esta situación no es usual, y desde luego, carece de importancia en el proceso interno de funcionamiento del RBC. Seria un caso extremo de entrada en el que el RBC tendría poco que aportar.

A pesar de la gran variedad de sistemas que existen, todos siguen unas pautas parecidas. Y el proceso de razonamiento que en general siguen todos ellos se puede descomponer en una serie de pasos bien conocidos con el nombre de *ciclo del RBC*. Este ciclo consta generalmente de cuatro etapas, ver Figura 2.2, las cuatro REs, (en inglés *Retrieve*, *Reuse*, *Revise* y *Retain*) [3];

1. *Recuperación* de casos similares a la descripción del problema.
2. *Adaptación* de la solución sugerida por el caso o casos más similares, de forma que se ajuste mejor al problema.

3. *Revisión* y evaluación del resultado obtenido al aplicar la solución seleccionada.

4. *Aprendizaje* de la nueva solución si ésta ha resuelto el problema con éxito o incluso también guardar o indexar la causa del error.

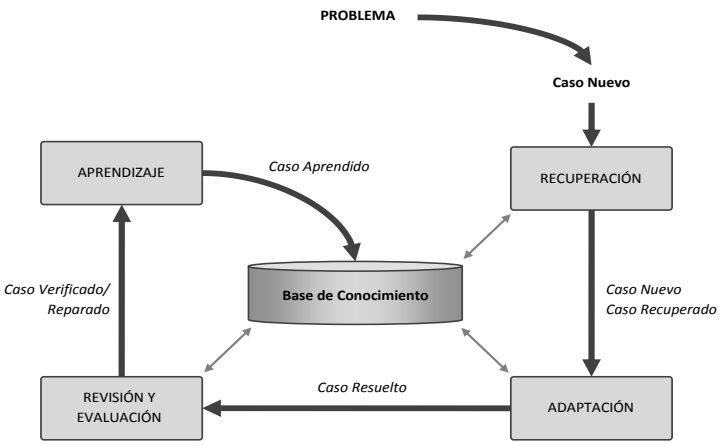


Figura 2.2: Ciclo de un sistema RBC

En los apartados siguientes se describe qué ocurre durante cada una de las etapas que componen el ciclo de un sistema RBC.

2.2.2.1. Representación e Indexación de los casos

Formalmente, un *caso* es una porción de conocimiento que representa el recuerdo que se tiene sobre una experiencia previa. Esta información dependerá del contexto en que sucedió. Por tanto, si se usa el *caso* para trabajar con RBC será importante guardar el contexto en el que se produjo la experiencia, ya que servirá para determinar cuándo es aplicable ese conocimiento. Los *casos* pueden tener tamaños diferentes, y son almacenados en una memoria llamada *base de casos*. Para crear una base de casos hay que tener en cuenta lo siguiente:

- La estructura y la representación de los casos.
- El modelo de memoria usado para organizar la base de casos.
- La selección de índices usados para identificar cada caso, si es que los casos fueron indexados.

Tanto la forma en que se represente un caso, como la manera en que se elijan sus índices, son partes importantes del proceso.

Representación de los casos

La *representación de los casos* es la forma en que se guarda la información que se le proporciona al sistema. Esta información se guarda o de forma exacta, tal y como llega al sistema, o tras realizar algún proceso de generalización entre los casos que ya había en memoria a lo que se conoce como *prototipo*. Un prototipo, requiere menor cantidad de memoria puesto que solamente se almacenará un caso cuando no se asemeje a otro ya existente. La Figura 2.3 [187] muestra las diferencias entre un prototipo y un caso que guarda toda la información. Normalmente, en una base de casos se compaginan ambos tipos de casos.

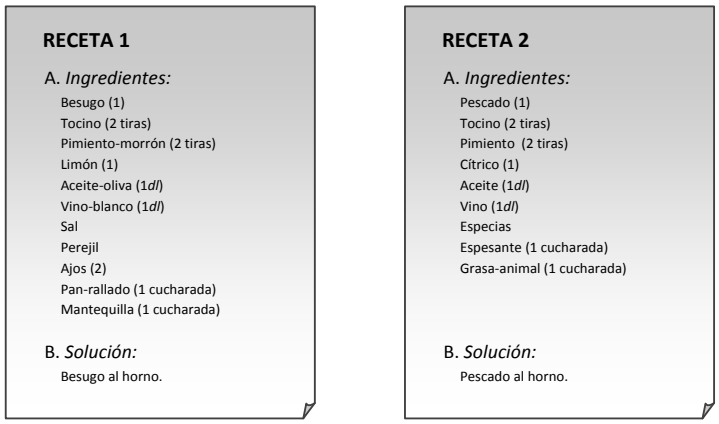


Figura 2.3: Almacenamiento de caso en forma de prototipo RECETA 2 versus almacenamiento de caso en forma exacta RECETA 1

En general, se dice que un *caso* queda representado, cuando de él se conocen las características siguientes:

- La descripción del problema, es decir, los objetivos y/o restricciones a satisfacer.
- Las soluciones del problema.
- Los datos iniciales.

- Una traza que indique como se ha llevado a cabo la resolución del caso, si fuera necesario.

Tradicionalmente los métodos de representación de casos se clasifican en tres categorías principales [17]: atributo valor, estructuradas y textuales. En la literatura existen otras más sofisticadas como muestran los siguientes trabajos [154, 183, 175, 81] que usan representación jerárquica o generalización de casos por citar algunas. Precisamente ésta es una de las características importantes del RBC que permite almacenar las experiencias anteriores tanto de forma estructurada, si el problema lo requiere, como desestructuradas, por eso es de aplicación prácticamente a cualquier tipo de problema. Nuestro estudio se hará sobre problemas cuyas bases de casos están estructuradas en la forma atributo-valor, si bien creemos que una extensión a otros tipo de estructura es posible.

Indexación de los casos

La *indexación* es una etapa opcional del proceso, pero cuando la librería de casos es grande, entonces la búsqueda de casos similares será muy lenta si se hace de forma secuencial. Para resolver este problema se acude a la *indexación*. Ésta, consiste en asignar uno o varios índices a cada caso para que así sea más fácil la recuperación del mismo y, además, se haga en un tiempo razonable. De ahí que sea muy importante la elección que se haga. Generalmente, cuando se va a elegir un índice, se tienen en cuenta los siguientes factores:

- Los índices deben ser lo suficientemente abstractos para que ayuden a recuperar el caso más adecuado, pero manteniendo un equilibrio, ya que si no, pueden hacer que el caso sea recuperado en demasiadas situaciones.

- Deben ser predictivos y reflejar la importancia de las características del caso y de los atributos.

- Deben describir los contextos en los que será adecuado recuperar el caso, por esto, los índices formaran parte de la información almacenada con el caso.

Aunque asignar índices es un proceso largo y manual que en ocasiones depende del experto, en la actualidad existen diversos estudios que indican cómo realizar la elección del índice más adecuado [19, 37, 86, 35, 170].

2.2.2.2. Recuperación de casos

La etapa de *recuperación* es el proceso de buscar en la base de casos, aquellos que sean lo más parecidos al caso actual. Es crucial en el desarrollo del sistema, puesto que si el caso recuperado no es adecuado para resolver el problema, todo el proceso se verá seriamente afectado. Frecuentemente, los casos a recuperar se buscan por casos enteros, comparando sus características con las del problema considerado, aunque hay veces en que es suficiente con sólo comparar una porción del caso. Este tipo de situaciones normalmente surgen cuando en la base de casos no existe un caso que coincida completamente con el caso que se está estudiando. El proceso de recuperar un caso es altamente dependiente, tanto del modelo de memoria, como de la forma en que los casos están indexados. Los métodos de recuperación empleados por los investigadores son muy variados y van desde el ya conocido método del Vecino más Cercano [44] hasta el uso de otros sistemas más complejos. Una discusión más detallada sobre estos métodos puede encontrarse en la Sección 2.3 de este capítulo. Pese a la gran variedad de métodos de recuperación que existen, hay ciertos factores que siempre es necesario tener en cuenta a la hora de aplicar uno u otro.

- El número de casos que contiene la base de casos.
- La cantidad de conocimiento disponible.
- La facilidad con la que se asignen los pesos a las características de cada caso.
- Si todos los casos deben o no ser indexados bajo las mismas características, es decir, si tienen o no atributos que varíen en importancia.

Una vez se ha seleccionado un método de recuperación y se ha recuperado un caso de la base de casos, se realiza un análisis para determinar si ese caso es suficientemente parecido al problema, o si la búsqueda debe ser modificada y reconducida nuevamente. Si durante este análisis, resulta que la elección del caso ha sido buena, ya habría finalizado esta etapa con un considerable ahorro de tiempo en las siguientes. Pero incluso habiéndose recuperado un caso apropiado, habría que tener en cuenta otro tipo de factores que nos harían considerar nuevamente la bondad del proceso de recuperación, como podrían ser, por ejemplo, el coste de la adaptación del caso recuperado, o las consecuencias que conlleva el aplicar esa solución adaptada (objetivo de nuestro estudio), entre otros.

2.2.2.3. Adaptación y Evaluación de los casos

La etapa de *adaptación* es el proceso de transformar la solución recuperada, en una solución apropiada para resolver el problema. La adaptación es una etapa importante dentro del ciclo de RBC, ya que es esto lo que añade la inteligencia a lo que de otra manera sería un simple modelo de emparejamientos. Hay dos formas generales de hacer la adaptación:

- Mediante la sustitución de aquellos valores que aparecen en el caso recuperado de la memoria, por aquellos otros valores del caso actual, de forma que la nueva solución hace uso de la situación actual que se quiere resolver.
- Mediante la aplicación, al caso que hay que resolver, del mismo conjunto de procedimientos, reglas o inferencias utilizados como solución del caso recuperado, el caso actual a resolver.

Si en el proceso de recuperación se obtiene un conjunto de casos candidatos recuperados, se puede o bien, seleccionar uno de ellos con algún criterio específico, o bien, hacer uso de las soluciones conjuntas de todos ellos en la búsqueda de una solución apropiada para el caso actual.

Como muestran los siguientes trabajos [73, 129, 67, 119, 97], existe una gran variedad de estrategias a seguir cuando se pretende adaptar un caso. Una vez elegida una posible estrategia y antes de que la etapa de adaptación se haya completado hay que comprobar si la solución adaptada tiene en cuenta las diferencias entre el caso recuperado y el problema actual, es decir, si realmente resolverá el problema. En este punto también hay que decidir qué hacer si la solución propuesta no resuelve el problema con éxito.

Posteriormente a la adaptación se realizará la evaluación del caso, o bien por un experto, o de forma semiautomática. Una vez evaluado se procede a la reparación si fuera necesario.

2.2.2.4. Aprendizaje y Mantenimiento de la base

Algunos razonadores basados en casos pueden carecer de la etapa de aprendizaje, o bien limitarla al simple almacenamiento de casos, si bien, es muy importante desarrollar esta etapa adecuadamente, puesto que dotará al RBC de capacidad inteligente.

Aprendizaje

Una vez se ha generado una solución apropiada para resolver el caso, ésta se prueba en el mundo real, analizando los resultados obtenidos y comprobando si resultan diferentes de lo esperado. Para ello, el sistema necesita de algún criterio capaz de valorar el rendimiento de la prueba. En la literatura existen numerosos métodos para tal fin, como puede verse en [98, 141]. Luego, el sistema se actualiza con la información obtenida acerca de la solución y éste será el medio del que el sistema se valdrá para aprender. Finalmente, esta nueva información se añadirá al sistema con dos propósitos: primero, aumentar la probabilidad de encontrar el caso más similar, y segundo, mejorar la solución que el sistema es capaz de crear.

Teniendo todo este en cuenta, el aprendizaje puede ocurrir no sólo cuando el caso ha sido solucionado con éxito sino también cuando no lo ha sido e interesa almacenar las causas del error. Por lo que resulta interesante tomar en cuenta ciertas consideraciones a la hora de añadir un caso a la base de casos:

1. *¿En qué situaciones debería el caso ser añadido a la base y en qué situaciones descartado?* Para decidir esto hay que tener en cuenta los siguientes factores:

   a) El nivel de éxito o fracaso de la solución

   b) Cómo de similar es ese caso a otro que ya esté en la base

   c) Si ese caso proporciona información relevante

2. *Si un caso es agregado a la base de casos, debe determinarse cómo añadirlo y como asignarle los índices, en el caso de que éstos sean necesarios para el problema.* Porque si el método de recuperación y los índices siguen una estructura muy rígida, la incorporación de un nuevo caso puede requerir una mayor planificación y reestructuración de la base de casos.

Mantenimiento de la base de casos

El mantenimiento de la base de casos es otra tarea que influye en el éxito del sistema. Por ejemplo, si la base crece sin orden podría llevar a resolver problemas de forma errónea, a almacenar casos conflictivos, a existir información redundante, etc. Una descripción interesante sobre un posible tipo de mantenimiento puede consultarse en [99], dónde para mantener de forma óptima la base de casos se implementan

políticas para revisar la organización de contenidos de la misma y así facilitar futuros razonamientos.

El mantenimiento de la base de casos generalmente envuelve, la tarea de añadir, revisar o eliminar casos, aunque también incluye cambios en el conocimiento. Cuando se utiliza RBC para resolver problemas, existe relación entre el número de casos almacenados en la memoria y lo eficiente que sea la recuperación. En principio, cuanto mayor es la librería de casos, mayor será el número de problemas cubiertos; sin embargo, también podría bajar la calidad del desarrollo del sistema el hecho de que el número de casos se hiciera demasiado grande. Por tanto, eliminar casos redundantes o casos que se consideren poco útiles para lograr un aceptable nivel de error, es una tarea importante para mantener con éxito un sistema RBC.

Para un correcto mantenimiento de la base de casos se han desarrollado muy diversos métodos. En la literatura pueden encontrarse trabajos que profundizan en ellos, como puede verse en [131, 109, 189, 100, 155].

2.2.3. Ventajas e inconvenientes de usar RBC

El RBC proporciona múltiples ventajas, entre las que cabe destacar:

1. Puede proponer soluciones de forma rápida sin tener que crearlas desde su inicio.
2. Reduce la tarea de adquisición de conocimiento.
3. Evita repetir errores cometidos en el pasado.
4. Puede proponer soluciones en dominios complejos donde no hay un modelo claro de representación del conocimiento.
5. Permite razonar con conocimiento incierto o impreciso.
6. Razona de forma parecida al ser humano, por lo que resulta más fácil comprender las soluciones y justificaciones que proporciona.
7. Utiliza la experiencia, como hacen los expertos, para aprender, prevenir soluciones erróneas, etc.
8. Es capaz de responder a situaciones excepcionales y poco comunes.

9. El sistema es fácil de mantener una vez desarrollado. Aprende incorporando nuevos casos y crece de forma progresiva; es escalable. Además, permite que se especialice en el dominio en que trabaja.

10. Puede utilizarse como almacén de conocimiento de una organización, lo que permite difundir experiencias y conocimiento experto por toda la organización y entrenar al nuevo personal.

11. Parte de soluciones globales, por lo que no es necesario descomponer el problema en subproblemas, resolver los distintos subproblemas y después unir las soluciones parciales evitando los problemas que puedan surgir.

12. Los casos son útiles para interpretar conceptos que no están claramente definidos.

El Razonamiento Basado en Casos también recibe críticas de algunos autores. Los detractores de este modelo de razonamiento argumentan que acepta evidencia anecdótica como principio de operación, sin datos estadísticos relevantes acerca del problema, por ello no hay garantía de que la generalización sea correcta. Las principales desventajas que presenta son:

1. El razonador puede tener tendencia a utilizar los casos recuperados a ciegas, sin comprobar si esas soluciones son válidas en la nueva situación.

2. Los casos previos pueden influir demasiado al razonador.

3. El conjunto de datos puede ser escaso para el tratamiento de un problema.

4. A menudo no se recuperan los conjuntos de casos más apropiados para razonar, sobre todo cuando el sistema es inexperto.

2.3. Selección y recuperación de casos

La etapa de *recuperación*, es una etapa muy importante dentro del ciclo de un sistema de Razonamiento Basado en Casos. Debido a que si el caso recuperado no es el adecuado, no podrá resolverse el problema correctamente y el sistema nos conducirá a cometer un error. Por tanto, no resulta extraño, que este problema haya sido estudiado por muchos investigadores y que multitud de algoritmos y técnicas diferentes hayan sido desarrollados con el objetivo de recuperar el caso más adecuado de la base de casos, que permita resolver con éxito el problema.

Para tal fin, una pregunta clave que debe hacerse a la hora de construir cualquier sistema que utilice Razonamiento Basado en Casos es: ¿cuál es el caso más similar al caso actual de entre todos los que actualmente están almacenados en la base de casos? La respuesta va a tener una importancia crucial en la fase de recuperación, y por tanto, en el comportamiento de todo el sistema. Luego, la forma en que se represente la similitud juega también un papel muy importante, ya que la mayoría de los algoritmos de recuperación y selección de casos están fundamentados en esto. Puesto que esta investigación se centra en la etapa de recuperación a lo largo de esta sección se hará un repaso, tanto del concepto de similitud, como de las técnicas más relevantes en recuperación, con el objetivo de situar nuestra propuesta en un marco teórico concreto.

2.3.1. Similitud

Pese a que existan razones para no usar un concepto tan sencillo como el de similitud en el proceso de recuperación, los buenos resultados empíricos obtenidos han conseguido que muchos investigadores acaben aceptándola como una herramienta importante en el proceso. No obstante, puede decirse que existen dos líneas de pensamiento diferenciadas [82, 83, 113, 117], por un lado los que piensan que un concepto tan sencillo no puede ser utilizado como representación del pensamiento humano y por otro, los que opinan que sí contribuye a dicha representación del pensamiento.

Aún así, el uso de la similitud está cada vez más extendido. Fueron Watson y Marir [174] los primeros que enfatizaron la importancia de las funciones de similitud en RBC, dónde su cálculo se convierte en una cuestión de vital importancia dentro del proceso de recuperación, puesto que en cierta manera, de este cálculo dependerá el caso recuperado. Por tanto, es importante en nuestro contexto establecer una apropiada función de similitud. Esta función puede interpretarse y representarse de muy distintas formas [138]. Estas son las más usuales:

1. Un predicado binario $SIM(x, y) \subseteq U \times U$, representando "$x$ e y son similares"
2. Un predicado binario $DISSIM(x, y) \subseteq U \times U$, representando "$x$ e y no son similares"
3. Una relación ternaria $S(x, y, z) \subseteq U^3$, representando "y es al menos tan similar a x como z lo es a x"

4. Una relación cuaternaria $R(x, y, u, v) \subseteq U^4$, representando "$y$ es al menos tan similar a x como u lo es a v"

5. Una función $sim(x, y) : U \times U \longrightarrow [0, 1]$, midiendo el grado de similitud entre x e y.

6. Una función $d(x, y) : U \times U \longrightarrow \mathbb{R}$, midiendo la distancia entre x e y.

Todas representan similitud, pero algunas son más expresivas que otras (de hecho, han sido escritas de menor a mayor grado de expresividad). Por ejemplo, la relación $S(x, y, z)$ permite definir el concepto "y es más parecido a x", mientras que la relación $SIM(x, y)$ no. Llegados a este punto surgen las siguientes cuestiones básicas:

I) ¿Cómo axiomatizamos estos conceptos, es decir, cómo hacemos para conocer las leyes que los rigen?

II) ¿Cómo están estos conceptos correlados, es decir, cómo distinguir entre los básicos y los que se obtienen a partir de los demás?

Por tanto, y teniendo en cuenta el contexto en que nos movemos, será razonable definir una medida que cumpla los siguientes axiomas intuitivos:

1. La similitud entre dos conceptos A y B está relacionada con sus atributos y características, de forma que, cuantas más cosas en común tengan más similares serán.

2. La similitud entre dos conceptos, también estará relacionada con las diferencias entre ellos, de forma, que cuantas más cosas diferentes tengan menos similares serán.

3. La máxima similitud se alcanza cuando A y B son exactamente iguales.

Estos axiomas se corresponden con las propiedades matemáticas de simetría, transitividad y reflexión. Sobre qué propiedades deberían ser exigidas a una función de similitud, existen opiniones divididas, y aunque los principios básicos para diseñar una medida de similitud son efectividad y simplicidad, los axiomas que normalmente se establecen son:

1. *Reflexiva*: $sim(x, x) = 1$

2. *Simétrica*: $sim(x, y) = sim(y, x)$

La mayoría de medidas de similitud, casi siempre son simétricas, y pueden, en principio, ser reflexivas y transitivas, o no cumplir ninguna de estas propiedades. También suele ser deseable que devuelvan valores acotados, puesto que facilita la comparación.

Un marco teórico para construir de forma sistemática medidas de similitud para RBC puede encontrarse en los trabajos [123, 124, 20]. Aunque usualmente las medidas de similitud más utilizadas en Razonamiento Basado en Casos son: la distancia Euclídea normalizada por la importancia de cada atributo (los pesos), la distancia de Hamming [66], la medida de Tversky [167], la de Hunt [75], la similitud del coseno, el índice de Jaccard [77], etc. En [105] aparecen recogidas todas las medidas más usuales en RBC.

Sin embargo, en la literatura se han desarrollado todo tipo de medidas. Sobre las que hacen uso de la lógica difusa [194], hay una gran variedad de trabajos publicados. Que van desde estudios comparativos sobre el comportamiento de estas medidas y sus propiedades: [199, 126, 36, 107, 172], hasta trabajos cuyo objetivo es presentar nuevas medidas diseñadas específicamente para contextos concretos como: [184] que las usa para medir la aproximación lingüística entre conceptos, [144] para la búsqueda de documentos, [12] para calcular la similitud entre (to estimate the similarity among relational cases represented using featureterms). Otras medidas de similitud entre conjuntos difusos pueden encontrarse en [52, 116, 171, 162, 50].

Otro gran avance en este campo, fue el desarrollo de las *medidas híbridas*, puesto que permitieron resolver el problema del cálculo de la similitud cuando el caso tenía atributos de distinta naturaleza. En la literatura existen numerosos ejemplos: [181] medida que calcula la similitud entre casos cuyos atributos son continuos, discretos o de ambos tipos, [105] medida híbrida para casos que mezclan atributos crisp con atributos difusos, [192] medida en la que cada atributo de un caso representa un tipo de contenido distinto, [61] esta otra medida utiliza tanto las cuestiones como las respuestas a dichas cuestiones para calcular la similitud. En esta línea actualmente existen multitud de trabajos, algunos ejemplos son: [195, 148].

Como puede apreciarse, la similitud es un tema muy estudiado y para dar solución a este problema se han desarrollado gran cantidad de medidas y aplicado gran cantidad de técnicas. Por ejemplo la teoría de la posibilidad para calcular la similitud entre casos cuando estos representan tiempo [80], la distancia de Bayes [90], o téc-

nicas más sofisticadas como [121, 64] o [68, 59] por poner algunas, ya que resultaría difícil citarlas todas.

2.3.2. Técnicas de recuperación

"*Similares experiencias guían futuros razonamientos resolviendo problemas e incluso permitiendo el aprendizaje*", este es el principal supuesto en Razonamiento Basado en Casos. De ahí parten las bases del desarrollo de las distintas técnicas de recuperación que nos ayudan en la resolución de problemas. Igual que ocurría con el concepto de similitud, no es sencillo desarrollar un método de recuperación de casos, ya que existen muchos factores a considerar a la hora de elegir el criterio que mejor se ajuste al problema que tratamos. Algunos de ellos son: el tipo y número de casos a ser buscados, la cantidad disponible de conocimiento sobre el dominio, la asignación de los pesos, el tipo de índice, etc.

Por tanto, en función del tipo de problema al que se enfrente el sistema y de la información que de éste se tenga, se deberá decidir qué factores habría que tener en cuenta a la hora de diseñar el algoritmo de recuperación. En la literatura existen multitud de algoritmos de recuperación que utilizan todo tipo de herramientas de Inteligencia Artificial como algoritmos genéticos, redes neuronales, árboles de decisión, lógica difusa, etc. No se discutirán todos estos métodos, sino solo aquellos que se han considerado más representativos, con la intención de ver como esta investigación es una propuesta original y que aporta valor.

Entre los métodos clásicos cabe destacar el k-NN [44], basado en el famoso principio del *Vecino más Cercano*, (en inglés *Nearest Neighbor* (NN) [46]). Este es uno de los algoritmos más sencillos que se han desarrollado en este campo. Funciona clasificando el caso por mayoría de votos de sus vecinos, es decir, se le asigna la clase (solución) que sea más común entre sus k vecinos. Normalmente k es un entero positivo y toma valores pequeños. La forma en que se elige el valor de k es muy importante porque ese valor puede evitar empates en los votos. El éxito de este algoritmo, radica en que pese a su sencillez los resultados experimentales son buenos en términos de precisión y coste computacional. Por este motivo, existen en la literatura muchas variantes que van desde estrategias para aumentar su eficiencia [21, 57, 95] o velocidad [168, 173], hasta variantes del mismo usando todo tipo de técnicas. En su contra, en problemas complejos donde intervienen diferentes tipos de factores, pasa por alto información importante para el problema y esto puede con-

ducir a este algoritmo, a cometer importantes errores. También entre los métodos clásicos, se encuentra el conocido *Fish and Shrink* [145] otro algoritmo sencillo de recuperación, que trata esta etapa desde un punto de vista diferente, haciendo uso de la información que aporta el experto. Una vez que los casos han sido almacenados en una base de casos, dicho experto describe el problema y los aspectos que bajo su punto de vista son los más importantes a tener en cuenta a la hora de recuperar el caso. Esto hace que la medida de similitud sea dada de forma natural y acorde con el problema, por lo que el algoritmo explora la base de casos rápidamente y presenta los casos más similares; con la desventaja de que los casos excluidos serán los menos similares bajo el criterio del experto, lo que podrá, en ocasiones, limitar las posibilidades del sistema.

En los últimos años, y siempre, con el objetivo de mejorar el rendimiento y la calidad de los sistemas RBC se han desarrollado multitud de técnicas de recuperación. Entre ellas cabe destacar, por ejemplo, aquellas que hacen uso de las redes neuronales como [71] donde se usan con éxito y gracias a ellas se consigue mejorar la precisión total del sistema, obteniendo así buenos resultados en clasificación. Mientras que en [43] las mezclan con el uso de agentes para crear un sistema inteligente basado en casos para planificación, con la capacidad de ofrecer alternativas. Otro enfoque donde se mezcla RBC y redes neuronales con éxito puede verse en [136] donde se crean sistemas de apoyo al diagnóstico. Cuando se resuelve un problema, la red neuronal es usada para fabricar hipótesis y así guiar el módulo del razonador durante la búsqueda del caso más similar que de soporte a la hipótesis. Aún así podemos encontrar todo tipo de técnicas sofisticadas que hacen uso de las redes neuronales como en [76, 32, 115, 106]. Trabajos en los que se aprovecha su eficiencia para el desarrollo de la tarea de recuperar y seleccionar, precisamente cuando los datos son incompletos, tienen ruidos, son imprecisos o cuando los dominios del problema son muy complicados. A pesar de los buenos resultados, no todo son ventajas, por ejemplo, el uso de redes neuronales puede no ser práctico ya que esperar recuperar un solo caso de cientos de posibles clases de casos haría demasiado grande el tamaño de la red y, por tanto, el proceso de recuperación muy lento. Por otro lado también podría llegar a ser excesivamente complicado aprender acerca de los casos dentro de una estructura de conexiones. Este enfoque, como se ha podido ver en los trabajos mencionados, difiere del propuesto en esta investigación por las siguientes razones: por un lado no tiene en cuenta el tipo de problema al que se enfrenta, es decir, no hace distinción entre problemas con y sin riesgo, y además, tampoco toma en cuenta las consecuencias que conlleva la decisión de recuperar un caso u otro.

También a la hora de recuperar un caso, es determinante la selección del peso de cada característica o atributo, puesto que influye directamente sobre la eficiencia y la precisión del sistema. El valor del peso que a veces es asignado de forma subjetiva, resulta determinante a la hora de recuperar un caso u otro, pero no permite garantizar qué ocurrirá con las soluciones recuperadas. Sin embargo, como se verá en el próximo capítulo, nuestro modelo sí será capaz de guiar el sistema hacia las soluciones adecuadas, minimizando el efecto, que en ocasiones puede ser negativo, del peso sobre la decisión y consiguiendo una recuperación más realista. Pese a esto, son conocidos los efectos positivos de usar pesos en la recuperación, por lo que se han propuesto diferentes métodos de selección de pesos [137, 62, 153, 65, 39, 188, 179], y multitud de técnicas como Analytic Hierarchy Processing (AHP) [127, 29], árboles de decisión [51], lógica difusa [157] o incluso técnicas estadísticas [180], todos ellos con resultados positivos.

Otra técnica utilizada para mejorar el funcionamiento de los modelos de Razonamiento Basado en casos es usar algoritmos genéticos como puede verse en [31] trabajo donde se integra RBC con algoritmos genéticos obteniendo así un sistema híbrido que consigue mayor precisión. También en [74] mezclan ambas técnicas para obtener un óptimo aprendizaje y en otros trabajos como [78, 150, 79]. Aunque los algoritmos genéticos aportan ventajas al RBC, no permiten introducir en el sistema las consecuencias negativas o positivas asociadas a cada decisión que tome el sistema, lo que les diferencia de nuestro enfoque. Por otro lado, utilizar árboles de decisión para mejorar la recuperación de casos, es otro método muy extendido en la actualidad como puede verse en [133, 178] o en [92] donde presentan un nuevo método de recuperación usando el algoritmo de búsqueda K-tree. También está muy extendida la idea de mezclar árboles de decisión con otras técnicas de IA como lógica difusa [151] o cluster para reducir el tiempo en la recuperación cuando la base de casos es muy grande como se hace en [112].

Pese a la multitud de técnicas estudiadas, a la hora de desarrollar la metodología objeto de esta investigación, se han utilizado sistemas de inferencia en lógica difusa y técnicas estadísticas propias de la Teoría de la Decisión, con el objetivo de mejorar la calidad de los sistemas RBC. Con ellas, se pretende aumentar la precisión en la clasificación y se introduce en la recuperación del caso más similar la información que aporta al problema el considerar las consecuencias asociadas a cada decisión. La elección de la lógica difusa se debe a que tiene la capacidad de modelar la reali-

dad de forma natural y como los sistemas RBC manejan juicios, evalúan, razonan y toman decisiones, ésta constituye una herramienta muy potente en su construcción. Por tanto, para hacer más eficaz el sistema se incorpora en este estudio. Si bien, algunos RBC han incorporado lógica difusa en la recuperación, creando novedosas técnicas como en [111, 16, 34] o [33, 38, 160] con la intención de mejorar la precisión, también uniéndolas con otras como en [30] que utiliza lógica difusa, árboles de decisión y algoritmos genéticos (en [40] se utilizan clasificadores bayesianos difusos con RBC para mejorar los resultados en la recuperación e incluso colonias de hormigas mezcladas con lógica difusa [94]), ninguno de ellos busca el mismo objetivo que se plantea en esta memoria.

También se han usado técnicas estadísticas como parte de sistemas de razonamiento basado en casos, como por ejemplo [135] o [128] que propone un método de selección de casos llamado Statistical Case-based reasoning (SCBR), el cual recupera el número óptimo de vecinos basándose en un medida probabilística de similitud. En [164] se usa Teoría de la Decisión para poner una cota a la probabilidad de que un caso no sea clasificado correctamente, [45] describe un sistema de RBC llamado Risk Cost Adviser (RICAD) el cual hace uso de funciones estadísticas para encontrar la respuesta que ofrezca la mayor confianza, en el sentido en que sea la respuesta cuya probabilidad de fallo sea menor. En concreto RICAD usa un factor de confianza para elegir las soluciones. Otros estudios usan sistemas de inferencia junto con RBC [166, 93, 110]. También es relativamente reciente en RBC ligar la etapa de adaptación con la de recuperación, de forma que se adapta el caso a la nueva situación a la vez que se recupera el más similar. Por ejemplo, [156] propone una nueva técnica, llamada adaptation-guided retrieval (AGR), esta técnica lo que hace es mezclar la similitud usada en la recuperación con las necesidades que surgen durante la etapa de adaptación. Los autores lo que plantean es aumentar la medida de similitud en función del conocimiento extraído de la adaptación del caso teniendo en cuenta si este será fácilmente modificado o no para ajustarse al problema. Similares técnicas de recuperación son usadas en los siguientes trabajos [120, 70].

Finalmente y para concluir, comentar que la investigación en este campo es extensa. Por este motivo se han seleccionado los trabajos que, en nuestra opinión, se consideran más relacionados con esta propuesta con el objetivo de mostrar su originalidad y mejorar la debilidad que presentan cuando se enfrentan a problemas que conllevan cierto riesgo asociado en sus decisiones.

2.4. Resumen y conclusiones

A lo largo de este capítulo, se han introducido los conceptos previos que se consideran necesarios, para llegar a situar esta investigación dentro un marco teórico adecuado, y poder así apreciar su originalidad y ventajas. Pudiendo, de esta forma, llegar a comprender mejor las razones que llevaron a ella.

En el capítulo se ha visto como RBC, resulta una metodología interesante con la que resolver problemas en el campo de la Inteligencia Artificial, lo que justifica su rápido crecimiento y el gran avance en su desarrollo. Por este motivo, se ha estudiado su historia y funcionamiento, analizando en detalle el concepto de similitud y la etapa de recuperación, puesto que son las partes del sistema que están directamente relacionadas con esta investigación. Para comprender con más profundidad el problema que presenta desarrollar una técnica adecuada de recuperación, también se han estudiado las medidas de similitud más significativas. La recuperación juega un papel crucial en el desarrollo del sistema puesto que si el caso recuperado no es adecuado para resolver el problema todo el proceso se vería seriamente afectado. Partiendo de que uno de nuestros objetivos es mejorar la calidad de la etapa de recuperación se ha analizado el estado actual de la investigación en este campo haciendo un repaso de las técnicas más destacadas. En este repaso se ha podido observar que ninguna de ellas tiene en cuenta si el problema al que se enfrenta pertenece o no la clase *Problemas Con Riesgo*. Por lo que pierde la valiosa información que aporta el considerar las consecuencias de recuperar un caso u otro. Luego surge de la necesidad de ver como sacar provecho a esta información.

En resumen, se puede afirmar que RBC es una herramienta muy intuitiva con la que resolver problemas en IA, que permite aprendizaje y también combinar distintos métodos de razonamiento. Estas características la convirtieron en punto de partida en nuestra investigación.

Capítulo 3

Un Nuevo Concepto en Recuperación: la Información de Riesgo

Según el principio en que se basa el Razonamiento Basado en Casos, la semejanza entre experiencias es la que nos guía en la resolución de problemas. De ahí la importancia del método elegido para recuperar y evaluar la similitud de un caso, especialmente, cuando el problema que se pretende resolver pertenece a la clase de problemas en los que el riesgo es un concepto relevante en el proceso de recuperación. Por tanto, en este capítulo, se añadirá un paso adicional en la etapa de recuperación, llamado Adecuación [28, 24, 25]. De forma que ahora al recuperar un caso, se tendrá en cuenta, no solo el valor específico del atributo, sino también si la solución considerada es o no adecuada para resolver el problema en función del riesgo producido por la decisión final. Este riesgo se introduce como un nuevo concepto en el problema al que se llamará "Información de Riesgo", y que será estudiado en detalle a lo largo de este capítulo.

3.1. Introducción

Como se comentó en el capítulo 1, en la vida real se pueden diferenciar claramente dos tipos de problemas: *Problemas Con Riesgo*(PCR) y *Problemas Sin Riesgo*(PSR) (ver Figura 1.1). Problemas del primer tipo son aquellos, en los que tomar una mala

decisión sobre como resolverlos puede derivar en graves consecuencias, por ejemplo: diagnosticar mal una enfermedad a un paciente, decidir invertir en una compañía poco fiable, etc. Mientras que *Problemas Sin Riesgo*, serán aquellos problemas para los que estas consecuencias no existen o son mínimas, es decir, tan poco costosas, que no son significativas.

Para los problemas dónde el riesgo es muy bajo o no existe, es decir, para los problemas en los que no elegir la solución correcta no produce consecuencias importantes, se han desarrollado multitud de técnicas de recuperación. Una recopilación de estas técnicas puede verse en el capítulo 2, dichas técnicas, aunque utilizan muy diversas herramientas de IA, no tienen en cuenta el riesgo de tomar una u otra decisión en el proceso de recuperación del caso más apropiado, (lo que en este contexto significa que no consideran las consecuencias que conlleva decidir recuperar un determinado caso). Sin embargo, este riesgo, es una información determinante, y que debe ser considerada a la hora de resolverlo.

Para tener en cuenta este tipo de información se define un nuevo concepto llamado *Información de Riesgo* [28, 24, 25] que mide cómo de apropiada o adecuada es una solución para resolver un caso. Este concepto solo se utilizará para resolver problemas de la clase PCR. Antes de definir este concepto formalmente, véase el siguiente ejemplo.

Invertir en empresas de éxito, es una proposición de negocio interesante. El Razonamiento Basado en Casos es una herramienta útil en muchos dominios y en particular en este campo, como muestran los siguientes artículos [102, 159, 85, 122, 103]. En este contexto, suponemos que nos encontramos ante la siguiente situación: hay tres casos almacenados en memoria *Empresa A*, *Empresa B* y *Empresa C*. Cada caso representa a una empresa de la que se conocen las siguientes características o atributos: el número de años que lleva funcionando, el sector al que se dedica, el beneficio medio obtenido en los últimos cinco años, si cotiza en bolsa, el número de empleados y el cash-flow medio obtenido en los últimos cinco años. También se tiene información de casos pasados de otras empresas en las que, en función del valor de esas características, se invertió o no en ellas. Por tanto, Invertir y No-Invertir, serán las soluciones asociadas a cada caso. Aparece una nueva compañía, *Empresa D*, en la que se quiere invertir, pero sin saber si resultará rentable. El experto ha informado previamente de que no es una compañía fiable y que, por lo tanto, no será adecuado invertir en ella. Veamos que ocurre si para decidir se utiliza un sistema RBC. En la

Tabla 3.1: Base de casos

Atributos	**Empr. A**	**Empr. B**	**Empr. C**	**Empr. D**
Número de años	80	1.5	30	10
Sector	Banca	Telecom.	Distribución	Distribución
Beneficio medio	3.25 %	-2.15 %	1 %	1.5 %
Cotiza en Bolsa	Si	No	No	No
Número de empleados	40,000	150	3,500	2,500
Cash-flow medio	10 millones	-2 millones	12 millones	-1.3 millones
SOLUCIÓN	INVERTIR	NO-INVERTIR	INVERTIR	¿?

Tabla 3.1, puede verse en detalle el valor de cada atributo y si la compañía resulta o no adecuada para invertir.

Para ver qué caso recuperar y por tanto qué solución aplicar, se calcula la similitud entre los casos almacenados en memoria y la *Empresa D*, en dos pasos. Primero, se calcula la similitud local entre atributos y luego, la similitud global entre casos. Para ello, se puede elegir entre multitud de medidas como ya se vio en el capítulo 2. Se han elegido las siguientes medidas de similitud local, Ecuación 3.1 cuando el atributo es categórico y Ecuación 3.2 en otro caso:

$$sim(x_i^{Mem}, x_i^{Nue}) = \begin{cases} 1 & si \quad x_i^{Mem} = x_i^{Nue} \\ 0 & si \quad x_i^{Mem} \neq x_i^{Nue} \end{cases} \tag{3.1}$$

$$sim(x_i^{Mem}, x_i^{Nue}) = 1 - \frac{|x_i^{Mem} - x_i^{Nue}|}{x_i^{max} - x_i^{min}} \tag{3.2}$$

dónde x_i^{Mem} es el i-*ésimo* atributo del caso en memoria x_i^{Nue} es el i-*ésimo* atributo del caso nuevo y x_i^{max}, x_i^{min} son el máximo y el mínimo respectivamente, teniendo en cuenta los valores del caso nuevo, para el i-*ésimo* atributo. En el segundo paso, se calcula la similitud global, haciendo la media aritmética de las similitudes obtenidas entre los atributos como indica la Ecuación 3.3:

Tabla 3.2: Similitud local y similitud global de los casos en relación al caso nuevo (Empresa D)

	Años	**Sector**	**Benef.**	**Cotiza**	**Empleados**	**Cash-flow**	**Sim. Globlal**
Empr. A	0.1083	0	0.6759	0	0.0590	0.1928	0.1727
Empr. B	0.8917	0	0.3240	1	0.9410	0.9500	0.6844
Empr. C	0.7452	1	0.9074	1	0.9749	0.0500	0.7800

$$Sim(C^{Mem}, C^{Nue}) = \frac{\sum_{i=1}^{n} sim(x_i^{Mem}, x_i^{Nue})}{n} \tag{3.3}$$

dónde C^{Mem} es el caso en memoria, C^{Nue} es el problema objetivo y n es el número de atributos en cada caso. Para evitar posibles confusiones, se notará a la similitud global como *Sim* y a la similitud local como *sim*.

Aplicando las Ecuaciones 3.1, 3.2 y 3.3, se obtienen los resultados mostrados en la Tabla 3.2, donde puede verse que el caso más similar a *Empresa D* es *Empresa C*. Por tanto, si se aplica la solución del caso recuperado, habría que invertir en *Empresa D*, y esto sería un error, puesto que el experto aconsejó no invertir en una empresa de estas características, ya que lo más probable es que sólo reportara pérdidas. Esto ha ocurrido, en parte, porque la medida no ha tenido en cuenta otros factores del problema que también son importantes. Por tanto, se hará más precisa la medida, introduciendo más información. A partir de ahora se tendrá en cuenta no sólo cómo de parecidos son dos atributos, sino también, la importancia que cada atributo tiene dentro del problema. Por ejemplo el atributo, *Cash-flow medio de los últimos 5 años*, que informa sobre las ganancias absolutas medias de la empresa en los últimos 5 años, no tiene la misma importancia a la hora de decidir si invertir o no en una empresa que el *Número de empleados* que ésta tenga. Luego es necesario introducir la importancia que cada atributo tiene en el problema, se hará utilizando la *variable peso* y se notará ω_i. La elección de los pesos es un paso importante a la hora de recuperar un caso. En el ejemplo, los pesos fueron dados por el experto. La Tabla 3.3, muestra los pesos de los atributos considerados para el ejemplo.

Para usar los pesos de cada atributo se debe modificar la fórmula de la medida de similitud global, que ahora se calcula como la media ponderada entre la similitud local de cada atributo y su peso correspondiente, según la Ecuación 3.4.

Tabla 3.3: Pesos de los atributos

Atributos	ω_i
Número de años	0.52
Sector	0.42
Beneficio medio de los últimos 5 años	0.63
Cotiza en bolsa	0.20
Número de empleados	0.33
Cash-flow medio de los últimos 5 años	0.81

$$Sim(C^{Mem}, C^{Nue}) = \frac{\sum_{i=1}^{n} \omega_i \cdot sim(x_i^{Mem}, x_i^{Nue})}{\sum_{i=1}^{n} \omega_i} \tag{3.4}$$

Una vez modificada la medida, se calcula nuevamente la similitud global entre los casos almacenados en memoria y el caso nuevo *Empresa D*, obteniendo los siguientes resultados: *Sim(Empresa A, Empresa D)= 0.2260*, *Sim(Empresa B, Empresa D)= 0.6690* y *Sim(Empresa C, Empresa D)= 0.6671*. Como puede verse, la *Empresa B* gana en similitud a *Empresa A* y *Empresa C*. Así que la solución recuperada es *No-Invertir*, que además resulta ser la solución adecuada para resolver el problema. Por tanto, se puede decir que los pesos aportan una información útil sobre el problema, pero tienen la desventaja de hacer fluctuar con facilidad a la medida. Es decir, pequeñas variaciones en los valores de los pesos, sin cambiar su orden de importancia, pueden provocar cambios en los resultados totales. Veamos un ejemplo, la Tabla 3.4, muestra los nuevos pesos modificados.

Al usar los pesos modificados de la Tabla 3.4, ω_i^{Nue}, se obtienen los siguientes resultados: *Sim(Empresa A, Empresa D)= 0.2293*, *Sim(Empresa B, Empresa D)= 0.6119* y *Sim(Empresa C, Empresa D)= 0.6844*. *Empresa C*, ahora obtiene un valor mayor en la similitud que *Empresa A* y *Empresa B*. Por tanto, la solución recuperada en este caso sería, la solución asociada a *Empresa C* (Invertir), solución nada aconsejable según la opinión del experto. Por tanto, usando solo los pesos y las similitudes en este tipo de problemas no se encuentra fácilmente el caso más adecuado. Para ello, sería útil tener en cuenta otra información que aporta el problema, y esto

Tabla 3.4: Pesos modificados

Atributos	ω_i	ω_i^{Nue}
Número de años	0.52	0.51
Sector	0.42	0.47
Beneficio medio de los últimos 5 años	0.63	0.65
Cotiza en bolsa	0.20	0.20
Número de empleados	0.33	0.32
Cash-flow medio de los últimos 5 años	0.81	0.80

significa, tener también en cuenta las consecuencias de tomar una decisión concreta antes de decidir que solución elegir.

Véase con detalle de qué tipo de información se trata. Para ello, se considera el atributo *Cash-flow medio de los últimos 5 años* y se supone que éste tiene valor negativo. Si además, la solución del caso considerado en ese momento es Invertir, y se sabe que una empresa con cash-flow negativo en los últimos 5 años no resulta fiable para invertir, la recuperación de ésta solución tendría consecuencias negativas, puesto que nos aconseja invertir en una empresa nada rentable. Luego, se trata de un atributo con mucha importancia, de hecho tiene un peso muy alto, pero con un factor negativo hacia casos cuya solución asociada sea Invertir, es decir, con *riesgo alto* de que la solución recuperada no sea adecuada para resolver el problema. Por tanto, se debería tener en cuenta este factor conflictivo al tomar la decisión sobre qué caso recuperar, especialmente cuando se trate de problemas de la clase *Problemas Con Riesgo*, sobre todo, porque este factor influye en la decisión y una mala decisión en esta clase de problemas conlleva graves consecuencias. Si por el contrario el *Cash-flow medio de los últimos 5 años* de una empresa es positivo, entonces el riesgo de que la inversión sea poco rentable es bajo. Como puede verse, ahora se trata de un atributo con un peso alto y además un factor adicional y positivo a favor de esa solución. Por tanto, el mismo atributo que tiene un peso alto, puede tener un riesgo alto o bajo, dependiendo de la decisión que se tome. El peso que aunque puede aumentar o bajar la importancia de un atributo en la decisión, no tiene en cuenta las consecuencias que conlleva la decisión tomada. Luego resulta interesante y útil considerar esta información en el proceso de recuperación.

Igual sucede cuando un atributo tiene poco peso, el riesgo de tomar una u otra

solución puede ser alto o bajo, éste dependerá también del valor específico que tome el atributo y de la solución del caso recuperado. Por ejemplo, si se considera un atributo con peso bajo y por tanto, con poca influencia sobre el resultado final, como *Número de empleados*. Cuando la *Solución del caso en memoria = Invertir* y *Número de empleados = 2*, el riesgo de que la inversión en esa empresa no sea rentable es muy alto, principalmente, porque empresas con sólo dos empleados no parecen muy sólidas. Luego, en este caso, tenemos un atributo con peso bajo y un factor que influye de forma negativa en la decisión final, por tanto con riesgo alto.

Se ha ilustrado a través de estos ejemplos que el uso exclusivo de los pesos puede no ser del todo apropiado para una recuperación con éxito. Por tanto, se pretende mejorar la recuperación introduciendo el *riesgo* de cada atributo. Concepto completamente diferente e independiente del concepto de peso, cuyas diferencias serán explicadas en detalle en la Sección 3.3. Teniendo en cuenta todo lo expuesto, puede verse la necesidad de considerar otro paso más durante el proceso de recuperación después de aplicar las funciones de similitud, a este paso se llamará *adecuación* de la medida.

Por último y como conclusión, comentar que a lo largo de este capítulo, lo que se pretende es proponer una nueva etapa dentro de la recuperación en RBC. Una etapa, que haga uso de la información proporcionada por el concepto de riesgo. Este concepto será definido como un conjunto de aplicaciones llamadas *Información de Riesgo Local* (IRL). Estas aplicaciones se utilizaran para incorporar esta importante información en el problema, lo que ayudará a recuperar el caso más conveniente o adecuado con el objetivo de conseguir una recuperación más eficaz.

El resto de este capítulo queda organizado en 4 secciones más: en la Sección 2 se define formalmente la *Información de Riesgo Local* como un conjunto de aplicaciones que se utilizarán para calcular la adecuación del caso. En la Sección 3, se estudian las diferencias entre riesgo y peso de un atributo, y además se observa cómo no es posible conseguir el mismo efecto en la recuperación usando solo los pesos que usando la *Información de Riesgo*. En la Sección 4 se evalúan los resultados obtenidos por el modelo, al que se llamará *Riesgo-RBC* (R-RBC), medidos en términos de precisión. Para ello, se utilizan bases de datos públicas procedentes del UCI-*machine learning repository* [60] y se comparan con los resultados obtenidos por otros modelos. Por último las conclusiones del capítulo serán expuestas en la Sección 5.

3.2. La adecuación sustituye a la similitud durante la recuperación

En esta sección, se introduce un nuevo concepto encargado de proporcionar una nueva información al problema, que se llamará *Información de Riesgo Local* [28, 24, 25]. Este concepto, medirá el riesgo que se corre, cuando la solución del caso en memoria se aplica para resolver el problema. Esto se hará atributo por atributo, y permitirá usar la valiosa información que se obtiene al descubrir si la solución del caso en memoria es o no apropiada para resolver el problema. Por tanto, esta nueva información ayudará a seleccionar el caso más conveniente o adecuado, el cual, en ocasiones, puede ser diferente del caso más similar como ocurría en el ejemplo presentado en la Introducción. Este ejemplo muestra como, cuando se busca el caso más similar en problemas de la clase PCR y para ello solo se consideran la importancia que cada atributo tiene en el problema y la similitud entre ellos, no siempre se consigue recuperar el caso más adecuado para resolver el problema. Para evitar esto y poder incorporar la información anteriormente mencionada, se introducirá este nuevo concepto.

Sea P un problema y $B = \{C_1, \ldots, C_q\}$ la base de casos asociada a ese problema, donde q es el número de casos. Cada caso, C, está formado por n atributos $C = \{A_1 = a_1, \ldots, A_n = a_n\}$, se debe tener en cuenta que el estudio se realizará sobre casos estructurados; $V_a = \{v_{a_1}, \ldots, v_{a_n}\}$, es el conjunto de valores donde cada atributo concreto toma su valor; $\Theta = \{S_j\}$, $j = 1, \ldots, m$ el conjunto de todas las soluciones asociadas al problema P y que están almacenadas en la base B y Ω el conjunto de etiquetas lingüísticas definidas para el riesgo. En este trabajo se ha definido $\Omega = \{bajo, medio, alto\}$.

Definición 3.1. Sea A un atributo, se define para cada valor de ese atributo el riesgo, R, como la aplicación $R : v_a \times \Theta \rightarrow \Omega$ que asigna la cantidad de riesgo asociada a cada par de valores $(A = a \in v_a, S_j \in \Theta)$.

Esta definición indica el riesgo que conllevaría aplicar la solución S_j para resolver el problema P, cuando el atributo A toma el valor concreto a de su rango de valores v_a.

Definición 3.2. Se llamara, *Información de Riesgo Local* (IRL) o simplemente *Información de Riesgo*, al conjunto de aplicaciones $\{R_i\}$, $i = 1, \ldots, n$, donde n es el

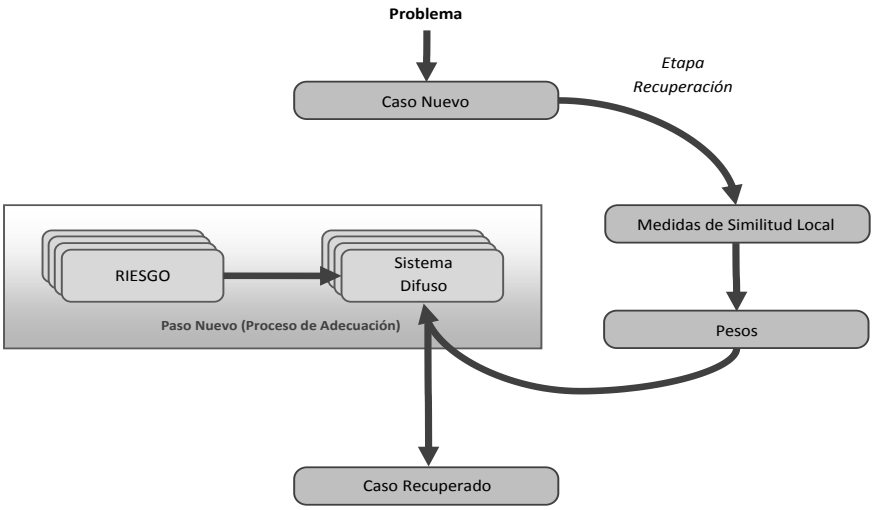


Figura 3.1: Proceso de recuperación considerando la Información de Riesgo en el problema

número de atributos de cada caso.

Por tanto, la *Información de Riesgo* es el conjunto de riesgos asociados a un caso. Si P es un problema que pertenece a la clase *Problemas Con Riesgo*, se debe considerar a la hora de resolverlo la *Información de Riesgo* de cada atributo $(R_1, \ldots, R_n)$. Esta nueva información será introducida en el proceso de recuperación a través de un sistema de inferencia difuso, como se explica más adelante. La Figura 3.1 muestra las diferencias entre el enfoque que se presenta en este trabajo y los métodos tradicionales.

Para introducir la IRL en la medida de similitud, se ha elegido un sistema de inferencia difuso: porque es útil y eficiente cuando se trata información imprecisa y porque es capaz de aportar a la medida tanto información lineal como no lineal procedente de las consecuencias de cada decisión. Para construir este sistema, primero se modifica la Ecuación 3.4 haciéndola depender del riesgo (R_i), del peso (ω_i) y de la similitud local $(sim(x_i^{Mem}, x_i^{Nue}))$, lo que permite obtener la *Adecuación* de cada caso en lugar de la similitud:

$$Adeq(C^{Mem}, C^{Nue}) = \frac{\sum_{i=1}^{n} f(R_i, \omega_i, sim(x_i^{Mem}, x_i^{Nue}))}{\sum_{i=1}^{n} \omega_i} \tag{3.5}$$

donde n es el número de atributos y la función $f(\cdot)$ es obtenida de forma implícita

a través del sistema de inferencia difuso y se calcula como la media ponderada de las salidas de todas las reglas disparadas. Usar la adecuación para recuperar un caso aporta las siguientes ventajas:

1. La recuperación ahora es más eficaz y realista, puesto que hace uso de la información local sobre cada atributo que aporta el riesgo.
2. Ahora la decisión es compartida entre el riesgo y el peso, eliminando la negativa sensibilidad que tiene el peso en la recuperación; por tanto, hace más robusta a la medida.

El sistema de inferencia difuso usado en la Ecuación 3.5 es una aplicación directa del modelo TSK [163]. Este sistema para el i-*ésimo* atributo contiene las siguientes 19 reglas:

Regla 1. Si R_i es *alto* y ω_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.1 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 2. Si R_i es *alto* y ω_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.2 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 3. Si R_i es *alto* y ω_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.3 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 4. Si R_i es *alto* y ω_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.2 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 5. Si R_i es *alto* y ω_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.3 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 6. Si R_i es *alto* y ω_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.4 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 7. Si R_i es *alto* y ω_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.3 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 8. Si R_i es *alto* y ω_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.4 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 9. Si R_i es *alto* y ω_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) - 0.5 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 10. Si R_i es *medio*, entonces $v_i = \omega_i \cdot sim(x_i^{Mem}, x_i^{Nue})$

Regla 11. Si R_i es *bajo* y ω_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.4 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 12. Si R_i es *bajo* y ω_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.3 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 13. Si R_i es *bajo* y ω_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.2 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 14. Si R_i es *bajo* y ω_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.2 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 15. Si R_i es *bajo* y ω_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.2 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$.

Regla 16. Si R_i es *bajo* y ω_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.1 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 17. Si R_i es *bajo* y ω_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.3 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 18. Si R_i es *bajo* y ω_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces $v_i{=}\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.1 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 19. Si R_i es *bajo* y ω_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces v_i=$\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}) + 0.1 \cdot (\omega_i \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Se ha elegido esta forma lingüística de expresar las reglas porque facilita la comprensión de las mismas y permite ver más fácilmente en qué situaciones la similitud local entre atributos crece o decrece en función del riesgo. Los términos lingüísticos (*bajo, medio, alto*) están definidos sobre R_i, ω_i y $sim(x_i^{Mem}, x_i^{Nue})$ respectivamente. Las etiquetas fueron definidas en consenso por los expertos que intervinieron en los experimentos, aunque en la literatura se pueden encontrar métodos que asignan dichas etiquetas [53, 22, 193]; por último, v_i $i = 1, \ldots, 19$ corresponde a la medida de similitud ajustada bajo la condición prescrita en la premisa de la i-*ésima* regla.

La construcción del sistema de reglas se hizo en dos pasos. Primero se diseñaron las reglas en función de la similitud, el peso y el riesgo, y luego se ajustaron los valores (0.1,0.2,0.3, ...) de los consecuentes de cada regla. Para ello, se siguió el siguiente procedimiento: primero se definieron como parámetros que tomaban distintos valores; de entre todos estos valores, con la ayuda de un experto y la experiencia del sistema, se fijaron aquellos con los que se obtuvieron mejores resultados. Veamos en detalle la razón del diseño de algunas de ellas. Por ejemplo, en la *Regla 1*, la similitud se rebaja una unidad, debido a que aunque hay riesgo alto, al ser la similitud y el peso también altos significa que se trata de un atributo importante y con un grado de similitud alto y no es aconsejable ignorar la influencia que esto tiene en el problema. Por tanto, se debe incluir el riesgo con un valor no muy elevado en esta situación. Sin embargo, en la *Regla 9* se disminuye la similitud en cinco unidades, puesto que en este caso se trata de un atributo con riesgo alto, poca importancia en el problema y con similitud baja con respecto al caso almacenado. Por otro lado en la *Regla 11* se aumenta en cuatro unidades la similitud, ya que es un atributo muy similar, con un peso alto y con un riesgo bajo, es decir, la solución considerada resulta adecuada para resolver problemas con este valor en el atributo. Por tanto, en este contexto un atributo con estas características debe tener más influencia en la recuperación. La *Regla 10* es una regla neutra, puesto que pensamos que el riesgo medio no es un valor importante como para afectar en la medida global.

Por último, la inferencia sobre el sistema difuso, o equivalentemente el cálculo de la *Adecuación*, se hace en dos pasos. Primero, se calcula la fuerza con que se dispara cada una de las reglas para el atributo i-*ésimo*, g_j^i, $j = 1, \ldots, k^i$, donde j indica la

regla disparada y k^i el número de reglas que se han disparado para el atributo i-*ésimo*. No se debe olvidar que para un mismo atributo se pueden disparar diferentes reglas. La fuerza con la que se disparan las k^i reglas, se calcula por agregación de todos los valores de las etiquetas lingüísticas evaluadas en su función de pertenencia utilizando el operador "y" (*and* en inglés), que se define como el producto algebraico. Por ejemplo, suponiendo que para el atributo i-*esimo* se ha disparado la *Regla 1*, la fuerza de disparo de esta regla g_1^i se calcula según la Ecuación 3.6

$$g_1^i = \mu_{alto}(R_i) \cdot \mu_{alto}(\omega_i) \cdot \mu_{alto}(sim(x_i^{Mem}, x_i^{Nue})) \tag{3.6}$$

Para todas las reglas salvo para la *Regla 10*, la fuerza de disparo se calcula usando la Ecuación 3.6. Para esta regla, por tratarse de una regla especial, la fuerza de disparo se calcula usando la Ecuación 3.7:

$$g_{10}^i = \mu_{medio}(R_i) \tag{3.7}$$

Una vez g_j^i se ha calculado para las k^i reglas disparadas para el atributo i-*ésimo*, se calcula la salida de la función $f(\cdot)$ como la media ponderada de las salidas de todas las reglas disparadas:

$$f(R_i, \omega_i, sim(x_i^{Mem}, x_i^{Nue})) = \frac{\sum_{j=1}^{k^i} v_j^i \cdot g_j^i}{\sum_{j=1}^{k^i} g_j^i} \tag{3.8}$$

donde k^i es el número de reglas disparadas para el i-*ésimo* atributo, v_j^i la salida de la regla j-*ésima* y g_j^i la fuerza con la que se dispara la regla j-*ésima* para el atributo i-*ésimo*. Un ejemplo sencillo de como calcular la fuerza de disparo de cada regla puede verse en el Apéndice A de esta memoria. Finalmente, la adecuación global del caso se calcula usando la Ecuación 3.5.

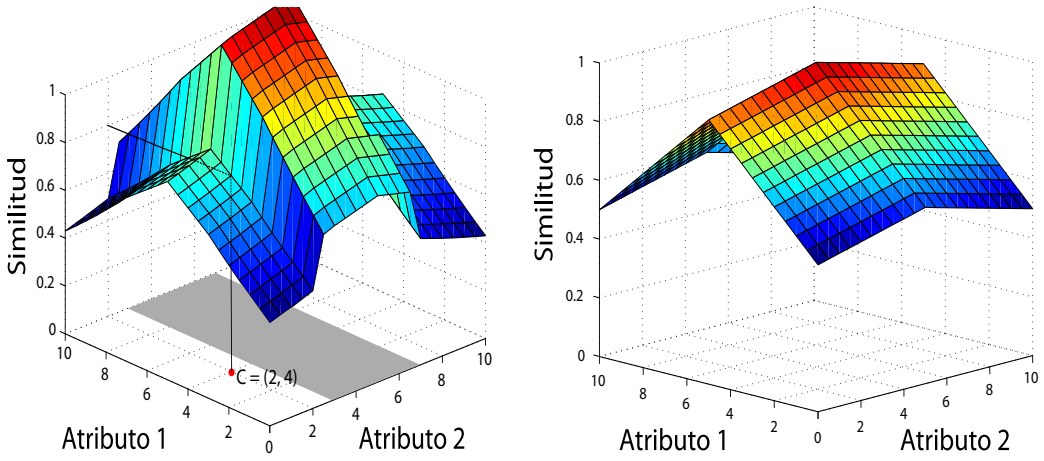
3.3. Diferencias entre peso y riesgo

Para evitar confusiones entre el concepto de peso de un atributo y el de riesgo asociado al valor del atributo, en esta sección se verá cómo la *Información de Riesgo* ayuda al sistema a recuperar el caso más conveniente con el que resolver el problema. También se verá por qué no es posible obtener esos resultados utilizando sólo los pesos de los atributos y las similitudes. Esto es debido a que el modelo que se obtiene con el ajuste de pesos no es el mismo que se obtiene usando la IRL en el problema. Un análisis gráfico de los dos modelos mostrará que se trata de dos enfoques muy

distintos e independientes en el sentido en que los resultados obtenidos con uno no pueden obtenerse con el otro.

Sea B una base de casos almacenada en memoria, donde cada caso, $C = (x, y)$, tiene dos atributos que se corresponden según la Figura 3.2 y la Figura 3.3 con los ejes de abscisas y ordenadas respectivamente. Los valores de cada atributo varían entre 0 y 10, como muestran dichas figuras. La intersección entre cada par de valores es un caso de la base de casos. Por ejemplo, el caso $C = (2, 4)$ se corresponde con el punto $(2, 4)$, como puede verse en la Figura 3.2(a). El eje Z, muestra los valores de la similitud y de la adecuación respectivamente en función del método. Un caso nuevo $C^{Nue} = (5, 5)$ llega al sistema y se desea conocer qué solución será la más conveniente para poder resolverlo con éxito. Para ello, se resuelve el problema usando un método de recuperación estándar representado por la función $f(\omega_i, sim(x_i^{Mem}, x_i^{Nue}))$, que no considera el riesgo y un modelo con riesgo $\widetilde{f}(R_i, \omega_i, sim(x_i^{Mem}, x_i^{Nue}))$. El objetivo es ver de forma gráfica cómo, pese a utilizar diferentes pesos, ambos modelos son distintos. Esto será debido principalmente al efecto que sobre la recuperación ejerce la función IRL. Para calcular la similitud local se usa la Ecuación 3.2 en ambos modelos. Mientras que para calcular la similitud global en el modelo de recuperación estándar se utiliza Ecuación 3.4, donde $C^{Nue} = (5, 5)$ y C^{Mem} todos los valores de la base de casos, ω_1 es 0.2, y ω_2 es 0.8. Para calcular la adecuación se utiliza el modelo de recuperación con riesgo, que se describió en detalle en la sección anterior y la Ecuación 3.5, en la que C^{Mem}, C^{New} y los pesos toman los mismos valores que en el modelo de recuperación estándar. Para asignar los valores de riesgo, se supuso que los casos cuya solución asociada era adecuada para resolver el problema eran los casos cuyos atributos tomaban valores entre 4 y 6, es decir, los casos situados en la zona sombreada de la Figura 3.2(a). Por lo que el riego, para los atributos cuyos valores estén en esa zona será bajo y para el resto de valores de los atributos el riesgo será medio. Los casos, según esto, más convenientes para resolver el problema serán aquellos cuyos atributos tomen valores entre 4 y 6, mientras que el caso más similar será el (5, 5).

En las Figuras 3.2(a) y 3.2(b), pueden verse las diferencias que presentan ambos modelos pese a usar los mismos datos, pesos y función de similitud local. La mayor diferencia está en los casos recuperados por cada método. Como se ve en las gráficas, el modelo con riesgo se ajusta mejor al problema y obtiene sus mejores resultados, justo en el área deseada, lo que significa que de no recuperar el caso más similar (5,5), al menos recuperaría un caso cuya solución es adecuada para resolverlo. Por



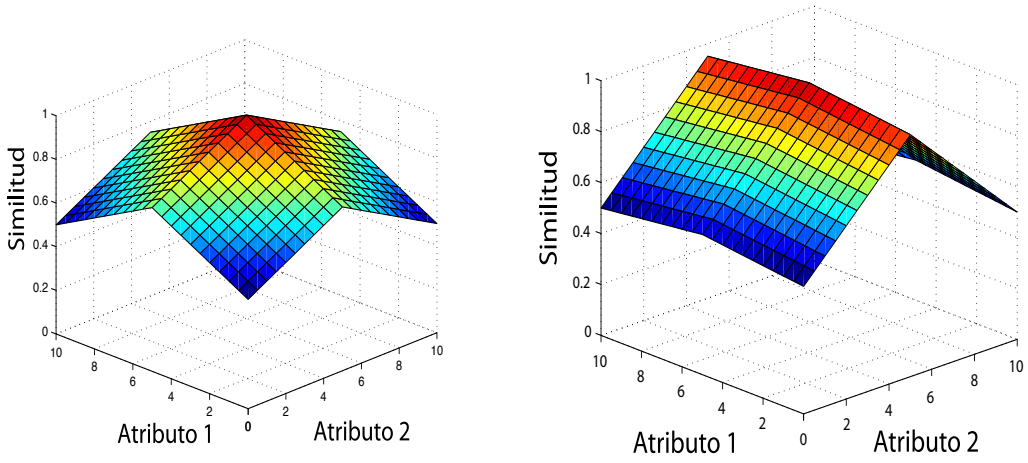
(a) Modelo con Riesgo($\omega_1 = 0.2$, $\omega_2 = 0.8$) (b) Modelo Tradicional($\omega_1 = 0.2$, $\omega_2 = 0.8$)

Figura 3.2: Análisis gráfico de los modelos I

tanto, mientras los métodos tradicionales (que son aquellos métodos basados fundamentalmente en similitud, como los descritos en el capítulo 2), obtienen similitud muy alta en un solo caso, el modelo con riesgo obtiene similitudes altas en la zona sombreada (zona que se corresponde con los casos cuya solución es adecuada para resolver el problema). Evitando problemas en la recuperación, puesto que la similitud es sensible a pequeños cambios.

Puede observarse también que los resultados obtenidos utilizando la *Información de Riesgo* para resolver el problema no pueden ser obtenidos, simplemente, cambiando los valores de los pesos. Es decir, no es posible concentrar los casos recuperados en el área deseada tan sólo cambiando los valores a los pesos. La Figura 3.3(a) y la Figura 3.3(b) muestran como se comporta un modelo tradicional frente a combinaciones diferentes de pesos $\omega_1 = 0.5$, $\omega_2 = 0.5$ y $\omega_1 = 0.9$, $\omega_2 = 0.1$.

Como puede verse en las gráficas, aunque cambia la situación del caso más similar, no se obtiene el mismo efecto que con el modelo de riesgo. Por ejemplo, en las Figuras 3.3(a) y 3.3(b) la similitud sin riesgo obtiene buenos resultados, con una concentración de la mayoría de casos recuperados en el área deseada, pero no es capaz de guiar o ajustar el proceso de recuperación hacia la zona de soluciones correctas como hace el riesgo. En conclusión, el modelo con riesgo dirige hacia una recuperación más adecuada evitando recuperar soluciones no deseadas.



(a) Modelo Tradicional($\omega_1 = 0.5, \omega_2 = 0.5$) (b) Modelo Tradicional($\omega_1 = 0.9, \omega_2 = 0.1$)

Figura 3.3: Análisis gráfico de los modelos II

3.4. Resultados experimentales

3.4.1. Descripción del problema: Casos de estudio

Para validar el sistema, se han usado cuatro bases de datos del UCI *machine-learning repository* [60]. Las bases de datos han sido seleccionadas con el objetivo de validar la utilidad de la información de riesgo en la recuperación de casos.

3.4.1.1. Los datos

BUPA liver disorder, fue recogida por BUPA Medical research Ltd. y donada para uso público por Richard S. Forsyth. En total está formada por 345 casos. Cada caso representa a un individuo de género masculino y cuenta con 6 atributos. Los cinco primeros son los resultados de los análisis de sangre realizados a cada individuo en los que se mide: el volumen medio del hígado (tamaño del hígado), la alcalina fosfotasa, la alamina aminotransferasa, el aspartate aminotransferase, y la gamma-glutamyl transpeptidase, el sexto atributo es la cantidad de alcohol que ingiere al día dicho individuo. La base está dividida en dos clases: *Clase 1*, corresponde a individuos sanos y *Clase 2* corresponde a los individuos que tienen problemas en el hígado derivados del consumo excesivo de alcohol. Solo se han utilizado 340 casos en los experimentos, para así construir subconjuntos de igual proporción y talla cara a la validación cruzada. Estos subconjuntos son seleccionados de forma aleatoria usando MATLAB 7.1. La Tabla 3.5 muestra los detalles de esta base de datos.

Tabla 3.5: Detalles de la base de datos usada en los experimentos (BUPA liver disorder)

Clase	Número de Casos	Porcentaje
Clase 1	140	41.18
Clase 2	200	58.82
Total	340	100.00

German Credit data, facilitada por el Dr. Hans Hofmann, Instituto de Estadística y Econometría, Universidad de Hamburgo (Alemania). Los datos provienen de las solicitudes de créditos realizadas por los clientes de un banco alemán. Contiene 1000 casos y cada caso está formado por 20 atributos. De ellos 7 están directamente relacionados con el préstamo: duración en meses del crédito, cantidad solicitada, cuota, número de años con el mismo lugar de residencia, edad, el número de créditos que el cliente tiene con ese banco y número de personas de las que es responsable. Y 13 indirectamente relacionados como: estado actual de la cuenta bancaria, historial de créditos, propósito para el que solicita el crédito, la cantidad de dinero ahorrado, años que lleva en su empleo, estado civil, si tiene o no garantías o aval, si posee propiedades, si el cliente tiene más bienes que paga a plazos, si la casa donde vive es o no propia, el tipo de trabajo que realiza, si tiene o no teléfono y por último si es o no extranjero. Estos casos están divididos en dos clases: *Clase 1*: indica que un cliente reúne las características para que le sea concedido el crédito y *Clase 2*: corresponde a clientes a los que no es conveniente que les sea concedido el crédito. Las clases contienen 700 y 300 solicitudes respectivamente. Se utilizan todos los casos para construir los subconjuntos que se usan en los experimentos. Como en la base de datos anterior los subconjuntos a usar en las validaciones cruzadas han sido generados de forma aleatoria con MATLAB 7.1. La Tabla 3.6 muestra los detalles de los subconjuntos usados en los experimentos.

Wine Recognition data, facilitada para uso público por el Instituto Farmacéutico, Análisis de Alimentos y Tecnologías de Italia. Esta base está formada por los datos originales obtenidos tras analizar químicamente distintas clases de vinos de una misma región italiana. Dichos vinos proceden de 3 cultivos diferentes que dan lugar a las tres clases en que se divide el conjunto de datos. En cada vino se analizaron 13

Tabla 3.6: Detalles de la base de datos usada en los experimentos (German Credit database).

Clase	Número de casos	Porcentaje
Clase 1	700	70.00
Clase 2	300	30.00
Total	1000	100.00

Tabla 3.7: Detalles de la base de datos usada en los experimentos (Wine Recognition database)

Clase	Número de casos	Porcentaje
Clase 1	50	31.25
Clase 2	70	43.75
Clase 3	40	25.00
Total	160	100.00

características o atributos, todas con valores continuos y que median: la cantidad de alcohol, el magnesio, la intensidad del color, etc. En total se analizaron 178 vinos de los que 58 pertenecían al primer cultivo, 71 al segundo cultivo y 48 al tercero. Solo se usaron 160 casos en los experimentos para poder así construir subconjuntos de igual proporción y talla cara a la validación cruzada. Los subconjuntos, como en los otros casos, han sido seleccionados de forma aleatoria usando MATLAB 7.1. La Tabla 3.7 muestra los detalles de cada subconjunto.

Glass Identification database, esta base ha sido donada por la Dra. Vina Spiehler y creada por B. German. En la escena de un crimen los cristales que quedan pueden ser usados como pruebas incriminatorias si son correctamente identificados. Esta es la razón por la que se estudió la clasificación de distintos tipos de cristales. El conjunto de datos está dividido en dos grupos principales: *Clase 1*, son cristales que provienen de ventanas de casas, coches, etc. y *Clase 2*, formada por el grupo de cristales con otra procedencia como vajillas, faros, etc. La base contiene 214 ejemplos de los que 163 pertenecen a la *Clase 1* y 51 a la *Clase 2*. Cada caso contiene 9 atributos: índice de refracción, la cantidad de sodio, la de magnesio, aluminio,

Tabla 3.8: Detalles de la base de datos usada en los experimentos (Glass Identification database)

Clase	Número de casos	Porcentaje
Clase 1	160	76.19
Clase 2	50	23.81
Total	210	100.00

silicio, potasio, calcio, bario e hierro. El procedimiento a seguir fue el mismo que en los ejemplo previos. La Tabla 3.8 muestra los detalles de cada conjunto de datos usado en los experimentos.

3.4.1.2. Ejecución de los experimentos

Se ha realizado validación cruzada de 10 subconjuntos para cada base de datos. En cada validación se toma la base de datos entera y se divide en 10 conjuntos excluyentes entre sí y con la misma distribución. Cada subconjunto se usa una vez como test para ver los resultados que se obtienen al probarlo contra el conjunto que resulta de unir los otros nueve restantes subconjuntos. Se hicieron 10 validaciones completas, lo que supone un total de 100 repeticiones de cada experimento, para tener datos suficientes y poder así contrastar los resultados con t-test (cuando las muestras de datos no son suficientemente grandes, es decir, menores de 30 se utiliza la teoría de muestras pequeñas para calcular la distribución [63, 101]).

Los resultados obtenidos en los experimentos son evaluados según la precisión. En este trabajo, la precisión mide la proporción de casos correctamente clasificados de entre todos los casos que han sido clasificados, es decir, la proporción de aciertos. El sistema, Riesgo-RBC (R-RBC), se implementó en MATLAB 7.1. Para contrastar su eficiencia se comparó con los siguientes modelos: J4.8, que es la implementación en WEKA del C4.5 [134] cuya versión comercial es el conocido C5.0), RBF Networks y k-NN con k=9 y k=3. De estos algoritmos se usó la implementación que de ellos está disponible en WEKA [182]. Para asegurar que los mismos subconjuntos que se generaron de forma aleatoria en MATLAB para la validación cruzada eran usados en WEKA, se trabajó con esta herramienta en línea de comandos, usando para ello las opciones t con la que se indicaba el conjunto de datos que debía coger y la opción T para indicar el conjunto de prueba. También se usaron en la comparación AN-

FIS (Adaptative Network Fuzzy Inference System) del que se usó la implementación disponible que hay en MATLAB 7.1 y RBC estándar implementado en MATLAB 7.1.

Por otro lado y antes de comenzar los experimentos, hubo que asignar los valores de peso y riesgo. Mientras que los pesos de cada atributo se aprendieron con WEKA, el riesgo de cada atributo fue asignado con ayuda de un experto. Gracias a las colaboraciones de nuestro grupo de investigación, fue posible contactar con expertos en todas estas áreas. Cada experto solamente tuvo que señalar las situaciones con riesgo alto y riesgo bajo, ya que como se ha comentado en este capítulo, el experto solo ha de asignar riesgo en situaciones críticas, lo que simplifica el trabajo. Veamos ahora como se hizo esta asignación en detalle. Para ello, se ha escogido aleatoriamente un caso de la base BUPA liver disorder, la Tabla 3.9 muestra los valores de sus atributos.

Tabla 3.9: Caso de muestra: BUPA liver disorders database

	mvc	**alkphos**	**sgtp**	**sgot**	**gammagt**	**drinks**
Valores	90	70	25	23	112	5
Pesos	0.27	0.48	0.85	0.98	0.67	0.27

El experto que asignó los riesgos en este experimento informó de que un valor mayor que 45 para un hombre en el atributo 5 (gamma-glutamyl transpeptidasa) indica claramente que su hígado debe estar afectado por un excesivo consumo de alcohol. Por tanto, el riesgo asociado a este valor de atributo será:

- Si *gamma-glutamyl transpeptidasa = 112* (ver Tabla 3.9) y *Solución del caso en memoria = Clase 1* (individuos sin ninguna alteración en el hígado), el riesgo de aplicar esta solución, es *alto*. Puesto que el experto dijo que ese valor en ese atributo era un claro indicativo de que existiesen alteraciones en el hígado, luego existe un riesgo alto al aplicar la solución del caso en memoria.

- Si *gamma-glutamyl transpeptidasa = 112* (ver Tabla 3.9) y *Solución del caso en memoria = Clase 2* (individuos con problemas en el hígado), el riesgo de aplicar esta solución es *bajo*. El experto informó de que un valor así en ese atributo indica hígado afectado, luego *Clase 2* resulta ser una solución adecuada para ese valor del atributo.

Veamos ahora qué ocurre con el atributo 6 (número bebidas alcohólicas por día):

- Si *número de bebidas = 5* (ver Tabla 3.9) y *Solución del caso en memoria = Clase 1* (individuos sin ninguna alteración en el hígado), el riesgo de aplicar esta solución, es *medio*.

- Si *número de bebidas = 5* (ver Tabla 3.9) y *Solución del caso en memoria = Clase 2* (individuos que presentan alteración en el hígado), el riesgo de aplicar esta solución, es *bajo*. El hígado de una persona que toma al día 5 bebidas alcohólicas tenga algún tipo de problema relacionado con esto.

Para el resto de atributos se sigue el mismo proceso con la ayuda del experto.

3.4.2. Resultados finales del estudio

En esta subsección se presentan los resultados obtenidos tras los experimentos. La precisión total de J4.8, RBF Networks, k-NN (k=*9*), k-NN (k=*3*), ANFIS, RBC estándar y Riesgo-RBC con cada uno de los conjuntos de prueba, puede verse en la Tabla 3.10, Tabla 3.11, Tabla 3.12 y Tabla 3.13. Bajo la columna *CV* aparece el número correspondiente a cada validación cruzada. La fila *DT* indica la desviación típica obtenida en cada método y la fila *Media* la media aritmética total. Observando las tablas puede verse que la precisión del modelo Riesgo-RBC, es la mayor entre todos los métodos en las pruebas con BUPA liver disorder, German credit y Glass identification. Sin embargo obtiene un tercer lugar en los experimentos con Wine recognition. Una de las razones por las que esto ocurre es que dicho problema no encaja totalmente en la clase *Problemas Con Riesgo* y es justo dentro de esta clase donde el modelo desarrollado obtiene los mejores resultados, puesto que ha sido diseñado especialmente para tratarlos.

Una vez resumidos los resultados obtenidos en los experimentos y para verificar que eran estadísticamente significativos, se contrastaron utilizando t-test con un nivel de confianza del 95 %. En la Tabla 3.15 puede verse el valor de los p-valores obtenidos al contrastar con ese nivel de confianza la media total de la precisión de Riesgo-RBC con la de cada uno de los otros métodos. Como puede verse Riesgo-RBC en el experimento con BUPA lier disorder presenta diferencias significativas respecto a la media con: J4.8, RBC, RBF Networks and k-NN con (k=*3,9*) a dicho nivel de confianza, salvo con ANFIS. También podemos ver como para la base German credit la precisión presenta diferencias significativas con respecto a todos los

Tabla 3.10: Precisión en los experimentos con BUPA Liver disorder

CV	**J4.8**	**ANFIS**	**RBC**	**R-RBC**	**RBF**	**k-NN(k=9)**	**k-NN(k=3)**
1	0.6558	0.6824	0.6265	0.7000	0.6617	0.6529	0.6382
2	0.6676	0.6971	0.6000	0.7177	0.6383	0.6147	0.6470
3	0.6617	0.7147	0.6147	0.7088	0.6500	0.6264	0.6235
4	0.6617	0.6852	0.6294	0.7205	0.6588	0.6264	0.6205
5	0.6323	0.7029	0.6264	0.7000	0.6529	0.6205	0.6147
6	0.6499	0.6794	0.6441	0.7235	0.6470	0.6529	0.6264
7	0.6823	0.6970	0.6382	0.7000	0.6411	0.6323	0.6264
8	0.6676	0.7147	0.6500	0.6911	0.6382	0.6353	0.6411
9	0.6147	0.7117	0.6588	0.7088	0.6588	0.6323	0.6205
10	0.6735	0.6882	0.6529	0.7029	0.6382	0.6264	0.6235
Media	*0.6567*	*0.6973*	*0.6341*	*0.7073*	*0.6485*	*0.6320*	*0.6282*
DT	0.0200	0.0133	0.0182	0.0093	0.0125	0.0103	0.0104

Tabla 3.11: Precisión en los experimentos con German Credit

CV	**J4.8**	**ANFIS**	**RBC**	**R-RBC**	**RBF**	**k-NN(k=9)**	**k-NN(k=3)**
1	0.7050	0.7130	0.6940	0.7540	0.7230	0.7370	0.7250
2	0.7110	0.7140	0.7050	0.7790	0.7420	0.7390	0.7240
3	0.7080	0.7140	0.7210	0.7440	0.7310	0.7360	0.7070
4	0.7150	0.7690	0.7060	0.7710	0.7300	0.7460	0.7180
5	0.7190	0.7400	0.7120	0.7670	0.7500	0.7470	0.7270
6	0.7150	0.7360	0.7200	0.7710	0.7340	0.7470	0.7210
7	0.6980	0.7410	0.7210	0.7590	0.7340	0.7360	0.7150
8	0.7110	0.7330	0.7120	0.7650	0.7420	0.7390	0.7200
9	0.7160	0.7160	0.7040	0.7570	0.7410	0.7300	0.7080
10	0.7000	0.7320	0.7150	0.7590	0.7360	0.7350	0.7230
Media	*0.7098*	*0.7308*	*0.7110*	*0.7626*	*0.7363*	*0.7392*	*0.7188*
DT	0.0070	0.0175	0.0088	0.0100	0.0076	0.0057	0.0068

demás métodos. Otro dato interesante es que pese a que es RBF quien obtiene mayor precisión en Wine recognition sus resultados no son estadísticamente significativos con respecto a los nuestros como puede verse en el valor del p-valor. En el último experimento (Glass identification) al mirar en Tabla 3.15 vemos que muestra diferencias significativas con todos los métodos salvo con RBC estándar y k-NN (k=*3*). Luego en general, y considerando el desarrollo y resultado de los experimentos, se puede afirmar que nuestro modelo con riesgo obtiene buenos resultados en precisión cuando trabaja con problemas en la clase *Problemas Con Riesgo*.

Tabla 3.12: Precisión en los experimentos con Wine Recognition

CV	**J4.8**	**ANFIS**	**RBC**	**R-RBC**	**RBF**	**k-NN(k=9)**	**k-NN(k=3)**
1	0.9125	0.8375	0.7875	0.9187	0.9750	0.9437	0.9437
2	0.8688	0.8500	0.8063	0.9125	0.9625	0.9500	0.9563
3	0.9187	0.8313	0.8063	0.9313	0.9875	0.9563	0.9375
4	0.9250	0.9688	0.9375	0.9500	0.9625	0.9375	0.9437
5	0.9313	0.9750	0.9500	0.9688	0.9750	0.9563	0.9500
6	0.9437	0.9688	0.9313	0.9625	0.9750	0.9437	0.9437
7	0.8938	0.9500	0.9187	0.9625	0.9750	0.9563	0.9500
8	0.9313	0.9812	0.9625	0.9625	0.9625	0.9563	0.9688
9	0.8875	0.9812	0.9563	0.9625	0.9625	0.9500	0.9625
10	0.9000	0.9812	0.9375	0.9688	0.9750	0.9435	0.9562
Media	*0.9113*	*0.9325*	*0.8994*	*0.9500*	*0.9712*	*0.9494*	*0.9512*
DT	0.0233	0.0649	0.0698	0.0097	0.0212	0.0084	0.0069

Tabla 3.13: Precisión en los experimentos con Glass Identification

CV	**J4.8**	**ANFIS**	**RBC**	**R-RBC**	**RBF**	**k-NN(k=9)**	**k-NN(k=3)**
1	0.9285	0.9238	0.9523	0.9619	0.9380	0.9142	0.8428
2	0.9285	0.9333	0.9619	0.9666	0.9333	0.9142	0.9571
3	0.9523	0.9380	0.9619	0.9523	0.9190	0.9285	0.9619
4	0.9333	0.9190	0.9571	0.9857	0.9333	0.9094	0.9523
5	0.9476	0.9142	0.9666	0.9666	0.9238	0.9142	0.9476
6	0.9380	0.9190	0.9619	0.9714	0.9285	0.9047	0.9571
7	0.9237	0.9142	0.9618	0.9666	0.9238	0.9142	0.9523
8	0.9333	0.9381	0.9476	0.9571	0.9190	0.9047	0.9380
9	0.9333	0.9333	0.9571	0.9666	0.9333	0.9142	0.9571
10	0.9333	0.9237	0.9571	0.9666	0.9285	0.9047	0.9428
Media	*0.9352*	*0.9257*	*0.9585*	*0.9661*	*0.9281*	*0.9123*	*0.9409*
DT	0.0087	0.0093	0.0055	0.0088	0.0065	0.0071	0.0352

Tabla 3.14: Ranking de la precisión obtenida por cada método

	J4.8	**ANFIS**	**RBC**	**R-RBC**	**RBF**	***k*-NN(*k*=9)**	***k*-NN(*k*=3)**
BLD	3	2	5	**1**	4	6	7
GC	7	4	6	**1**	3	2	5
WR	6	5	7	3	**1**	4	2
GI	4	6	2	**1**	5	7	3

Tabla 3.15: Resultados de los t-test al contratar Riesgo-RBC con cada modelo

	J4.8-RCBR	**ANFIS-RRBC**	**CBR-RRBC**	**RBF-RRBC**	***9*-NN-RBC**	***3*-NN-RRBC**
BLD	*0.000*	0.156	*0.000*	*0.000*	*0.000*	*0.000*
GC	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*	*0.000*
WR	*0.001*	0.356	0.034	0.851	*0.004*	0.580
GI	*0.000*	*0.000*	0.033	*0.000*	*0.000*	0.048

3.5. Resumen y conclusiones

Como ya se comentó en el capítulo anterior, la mayor parte de las técnicas de recuperación se limitan a buscar el caso más similar y no el caso más adecuado. Normalmente, estas técnicas recuperan un caso en función de su peso y similitud. En este capítulo se ha ampliado el proceso de recuperación de un caso, introduciendo un nuevo paso llamado *Adecuación*. Para calcular la adecuación se ha definido el concepto, *Información de Riesgo*, que mide lo adecuada que es una solución para resolver un problema en función del valor del atributo. Este concepto se define localmente y se asigna de manera independiente a cada atributo teniendo en cuenta su valor concreto y la solución considerada. Esta información, se introduce en el proceso de recuperación a través de un sistema de inferencia difuso. Integrar RBC con un sistema de inferencia difuso permite beneficiarse de las ventajas de estos sistemas, como son: su eficiencia cuando tratan información imprecisa y su capacidad para aportar a la medida información lineal y no lineal, entre otras. También permite complementar sus debilidades facilitando la recuperación de alternativas dentro de conjuntos grandes de casos. Usar la adecuación para recuperar un caso permite que la recuperación sea más realista y eficaz, y además, consigue que la decisión sea compartida entre el riesgo y el peso eliminando la sensibilidad negativa que tiene el peso en la recuperación y por tanto haciendo a la medida más robusta.

En este capítulo, también, se incluye un estudio gráfico detallado sobre las diferencias entre riesgo y peso de un atributo. En dicho estudio se puede comprobar como, variando solo los pesos, el sistema no es capaz de ajustar el proceso de recuperación y dirigirlo hacia la solución más adecuada. Por último, se aplica el modelo a cuatro conjuntos de datos, tomamos del UCI *machine-learning repository*, y así comprobar como se comporta. Los resultados obtenidos son comparados con otros métodos conocidos como: J4.8 (la implementación en Weka del conocido C4.5), RBF-Networks, ANFIS, k-NN (k=3, 9)y RBC convencional. Al comparar en términos de

precisión los resultado obtenidos, se pudo ver como el método propuesto mostraba los mejores resultados en media. Las diferencias entre estos resultados han sido contrastados con un nivel de confianza de 95 %. Esto prueba como el nuevo concepto de riesgo mejora la recuperación en problemas de la clase *Problemas Con Riesgo* y guía a los pesos hacia una mejor recuperación.

Capítulo 4

Modelos para Definir la Función Información de Riesgo

En el capítulo anterior, se presentó el concepto de Información de Riesgo, que permite considerar aspectos del problema que hasta ahora no habían sido tenidos en cuenta. Su principal desventaja, es la dependencia total del experto. Por tanto, en este capítulo, se plantean soluciones para cuando el experto da sólo parte de la información, o no se dispone de información alguna. Para ello, se propondrán métodos que resuelvan este problema, se probaran bajo distintas hipótesis que simularán situaciones reales y se hará un análisis de su comportamiento, para así poder extraer conclusiones de cuándo será adecuado aplicarlos. Esto permitirá utilizar la IR en una mayor variedad de situaciones sacando, por tanto, el máximo provecho posible de ella.

4.1. Introducción

En el capitulo anterior se estudió, cómo el considerar la información que el concepto de riesgo aporta al problema, permitía que el sistema obtuviese buenos resultados; aunque con el inconveniente de que era dada en su totalidad por el experto. Para poder utilizar esta información en un mayor abanico de situaciones, en este capítulo se definen y analizan modelos para calcular el valor de la función IR cuando la información del experto este incompleta, sea difícil de interpretar o imprecisa, inclusive abordamos el no contar con ninguna información. A la hora de organizar este estudio se han considerado estas tres situaciones:

1. Que un experto esté siempre a nuestra disposición.

2. Que el experto tan solo de información parcial del problema.

3. Que no se disponga de información alguna, es decir, cuando sea imposible, por las razones que sean, contactar con un experto en el tema.

La primera situación, en la que se cuenta plenamente con la ayuda del experto, es la misma situación del capítulo anterior. En este caso se aplicaría la metodología descrita en dicho capítulo. Esta situación, pese a ser la ideal, en la vida real suele ser muy poco frecuente, por ser demasiado costosa e incluso en ocasiones inviable, puesto que resulta ilógico tener el sistema dependiendo de una persona a lo largo de todo su ciclo de vida.

En la segunda situación, se plantea el interrogante de cómo actuar cuando sólo se dispone de parte de la información. Abstrayendo el problema se puede plantear el siguiente objetivo: "*conocida cierta información sobre una función desconocida f, proponer modelos con los que construir el candidato a f y analizar en cada situación cuál resulta más adecuado y mejor se ajusta al problema*". En la literatura existen numerosos métodos que abordan este problema. De entre todos, se han elegido algunos de los más comúnmente utilizados, con la intención de hacer un muestreo y analizar en qué tipo de problema y/o situación son adecuados. Se han seleccionado, de las distintas familias de métodos, los más representativos teniendo en cuenta los siguientes factores:

- *El Tipo de Información*: ¿cómo dan esa información?, ¿cómo la información que proviene del exterior puede ser expresada?, ¿en forma numérica?, ¿lingüística?, ¿a modo de propiedades?, etc.

- *La Cantidad de Información*: ¿de cuánta información se dispone?, ¿son muchos datos o muy pocos?, etc.

- *La Calidad de la Información*: ¿cómo de buena es la información?, ¿se trata de una información exacta o aproximada?, ¿es fácil interpretarla?, ¿proviene de datos muestrales, o directamente del experto?, etc.

Por ejemplo, se proponen diferentes modelos de *Interpolación*, cuando la información venga de datos muestrales o el experto advierta que ha de cumplirse con exactitud. En estos casos la información vendrá dada punto a punto, es decir, serán

valores fijos que se deben cumplir con rigor, y/o también ciertas propiedades que han de verificarse como el crecimiento o decrecimiento, concavidad o convexidad, etc. Sin embargo, se proponen modelos de *Aproximación* cuando se conozcan algunos valores de la función, pero que lo que interese sea que simplemente se aproxime el candidato a f, en el sentido que se considere conveniente a esos valores y/o que además cumpla alguna condición que se exija según las características del problema. Otra razón por la que se han elegido estos modelos es por ser buenos modelos matemáticos, rigurosos y muy adecuados en el caso que tan sólo interese "aproximarse" a la información que dio el experto. Puede ocurrir, además, que o bien lo que se tenga sea una descripción lingüística de la información o que dicha información esté mezclada con información numérica, en estos casos, se proponen modelos de aproximación mediante reglas difusas.

Finalmente si no se dispone de información alguna sobre la función de riesgo, ni con ayuda de un experto, se ha desarrollado un método propio [26] diseñado específicamente para aprender la función de riesgo a partir de los casos almacenados en memoria. Este método consiste en un modelo de asignación probabilística, el cual al no disponer de información del exterior, utiliza la información almacenada en la base de casos para asignar el valor de riesgo a cada atributo. También podrá ser de utilidad como complemento para alguno de los otros métodos.

El resto del capitulo se organiza de la siguiente manera: en la Sección 2, se exponen de forma detallada los métodos propuestos para aproximar la IR cuando el experto sólo da información parcial del riesgo. En la Sección 3, se estudia el caso de cómo aproximar la IR cuando no se tiene información alguna proveniente del exterior. Luego, un estudio comparativo de los distintos métodos se presenta en la Sección 4, comparándolos, tanto entre ellos, como con los resultados obtenidos con la ayuda del experto. Finalmente, un resumen junto con las conclusiones del capítulo serán expuestos en la Sección 5.

4.2. Asignación de riesgo en ausencia parcial del experto

A lo largo de esta sección se proponen y analizan distintos modelos para poder utilizar la IR pese a no disponer de toda la información necesaria. Se trata de proponer soluciones para las situaciones en que el experto no esté completamente a

nuestra disposición, que la información provenga de ejemplos o datos muestrales pero esté incompleta, etc. También es importante tener en cuenta que cada problema es diferente y que por tanto la información vendrá dada de diversos modos: en forma de datos numéricos aislados, de forma lingüística o aproximada, etc. Luego se pretende sacar el máximo partido a esta información, y además no perder nada o intentar perder la menos posible, haciendo uso de los métodos propuestos. Realmente, esta situación será la que más se presente en la práctica. Ahora se analizará cada método de forma detallada.

4.2.1. Interpolación

La interpolación es un conocido método matemático que propone soluciones al problema clásico de calcular el valor de una función en un punto, cuando éste no se puede evaluar directamente en la expresión de la función por las razones que sean. Esta técnica usa como criterio de aproximación a la función f que buscamos, la coincidencia en un número finito de datos con la función de partida (función que o bien no conocemos o que es difícil de manejar). La función a calcular, además, deberá pertenecer a un espacio vectorial prefijado, este espacio será el espacio de las funciones sencillas de trabajar, por lo que normalmente será el espacio de las funciones polinómicas o polinómicas a trozos.

Luego, a grandes rasgos resolver un problema de interpolación, consiste en encontrar una función, f, fácil de construir y a la vez fácil de evaluar, que coincida con exactitud con la función objeto del problema en los datos que se conocen sobre ésta. En estas condiciones, se dice que la función así construida interpola a la función dada en dichos datos. A la función obtenida se le denomina *interpolante*, *función interpoladora* o *función de interpolación*.

Como puede verse, esta metodología propone una solución posible para el problema planteado cuando la información es dada en forma de valores puntuales, y/o además son conocidas ciertas propiedades sobre crecimiento, decrecimiento, concavidad, etc. Partiendo de este tipo de información la Interpolación resulta una herramienta adecuada. Veamos ahora el problema de interpolación polinómica de forma abstracta:

Planteamiento general del Problema de Interpolación

Se entiende por *Problema General de Interpolación* (PGI) el que se plantea a continuación. Sea V un espacio vectorial de dimensión n sobre $\mathbb{R}$ y $L_1, \ldots, L_n$, n formas lineales sobre V, esto es, n aplicaciones lineales sobre V, $L_i : V \to \mathbb{R}$. Dados $z_1, \ldots, z_n \in \mathbb{R}$ resolver el PGI es encontrar un elemento $f \in V$ tal que:

$$L_i(f) = z_i, \quad \forall i = 1, \ldots, n$$

La solución a este problema viene dada por el siguiente teorema:

Teorema 4.1. *Existencia y unicidad de solución del PGI*
Las siguientes afirmaciones son equivalentes:

I) Existe un único elemento $f \in V$ tal que

$$L_i(f) = z_i, \quad \forall i = 1, \ldots, n$$

II) 0 es el único elemento de V tal que

$$L_i(f) = 0, \quad \forall i = 1, \ldots, n \tag{4.1}$$

III) Para cualquier base $\{f_1, \ldots, f_n\}$ de V se tiene que $det(L_i(f_i)) \neq 0$

IV) Existe, al menos, una base $\{f_1, \ldots, f_n\}$ de V tal que $det(L_i(f_i)) \neq 0$

V) $\{L_1, \ldots, L_n\}$ son linealmente independientes en V y por tanto son base de V.

Si existe la solución del PGI esta es única y puede escribirse como sigue. Dados $z_1, \ldots, z_n \in \mathbb{R}$ el único $f \in V$ tal que $L_i(f) = z_i$ se escribe en la forma:

$$L_i(\sum_{j=1}^{n} a_j f_j) = \sum_{j=1}^{n} a_j L_i(f_j) \quad i = 1, 2, \ldots, n$$

donde $a_1, a_2, \ldots, a_n$ son números reales.

Para construir la solución al PGI se han seleccionado los siguientes métodos de interpolación: Lagrange, Hermite e Interpolación mediante funciones splines, por ser los más usuales y también los que más se ajustan a las condiciones del problema. A continuación, se verá en detalle, cada uno de los métodos.

4.2.1.1. Interpolación de Lagrange

La interpolación polinómica en forma de Lagrange es uno de los métodos de interpolación más usuales y estudiados. Se ajusta a la perfección al problema planteado cuando la información de partida son valores puntuales que han de verificarse con exactitud, es decir, cuando se tengan puntos concretos que la función IR asociada a ese atributo ha de verificar con exactitud. El método de interpolación de Lagrange construye el polinomio que se aproxime a la función buscada, a partir de ciertos puntos del plano por los que debe pasar su gráfica, por tanto cuando en la práctica la información que de el experto sea de este tipo, éste método será adecuado para conocer el resto de valores de la función IR.

Determinar un polinomio de primer grado, que pase por dos puntos distintos, (x_0, y_0) y (x_1, y_1), es lo mismo que aproximar una función f para la cual $f(x_0) = y_0$ y $f(x_1) = y_1$ por medio de un polinomio de primer grado que interpola, o que coincide con los valores de f, en los puntos dados. Para construir estos polinomios mediante el método de Lagrange han de definirse las funciones L_0 y L_1:

$$L_0(x) = \frac{x - x_1}{x_0 - x_1} \; y \; L_1(x) = \frac{x - x_0}{x_1 - x_0}$$

Estas definiciones implican que:

$$L_0(x_0) = \frac{x_0 - x_1}{x_0 - x_1} = 1, \;\; L_0(x_1) = \frac{x_1 - x_1}{x_0 - x_1} = 0, \;\; L_1(x_0) = 0 \;\; y \;\; L_1(x_1) = 1$$

Se define ahora el polinomio de interpolación, $P(x)$, como:

$$P(x) = L_0(x)f(x_0) + L_1(x)f(x_1) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1),$$

con lo cual verifica

$$P(x_0) = 1 \cdot f(x_0) + 0 \cdot f(x_1) = f(x_0) = y_0$$

y

$$P(x_1) = 0 \cdot f(x_0) + 1 \cdot f(x_1) = f(x_1) = y_1$$

Es decir, P es la única función lineal que pasa por (x_0, y_0) y (x_1, y_1), como muestra la Figura 4.1.

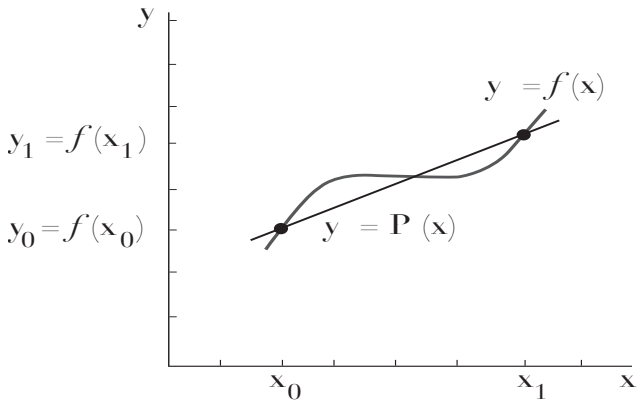


Figura 4.1: Función interpoladora de los puntos (x_0, y_0) y (x_1, y_1)

Para generalizar esto a polinomios de grado mayor se construirá, a modo de ejemplo, un polinomio de grado menor o igual que n que pase por los siguientes $n+1$ puntos $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))$ fijados (véase Figura 4.2). En este caso, se necesita construir, para cada $k = 0, 1, \ldots, n$, un polinomio de grado n, que se denota, $L_{n,k}(x)$, con la propiedad de que $L_{n,k}(x_i) = 0$ para $i \neq k$ y $L_{n,k}(x_k) = 1$. Para que se cumpla la primera propiedad $L_{n,k}(x_i) = 0$ para $i \neq k$, el numerador de $L_{n,k}(x)$ debe contener el producto:

$$(x - x_0)(x - x_1) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n)$$

Y para que se cumpla la segunda propiedad exigida, $L_{n,k}(x_k) = 1$, el denominador de $L_{n,k}(x)$ debe ser dicho producto evaluado en $x = x_k$. Es decir,

$$L_{n,k}(x) = \frac{(x - x_0) \ldots (x - x_{k-1})(x - x_{k+1}) \ldots (x - x_n)}{(x_k - x_0) \ldots (x_k - x_{k-1})(x_k - x_{k+1}) \ldots (x_k - x_n)}$$

A $L_{n,k}$ se le llama k-*ésimo coeficiente polinomial de grado n de Lagrange*. La Figura 4.3 muestra un esbozo de la gráfica de un $L_{n,k}$ típico. Una vez conocida la fórmula de $L_{n,k}$, el polinomio de interpolación se describe fácilmente:

Definición 4.1. Se define el *n-ésimo polinomio interpolador de Lagrange de la función* f como:

$$P_n(x) = f(x_0)L_{n,0}(x) + \ldots + f(x_n)L_{n,n}(x) = \sum_{k=0}^{n} f(x_k)L_{n,k}(x),$$

siendo

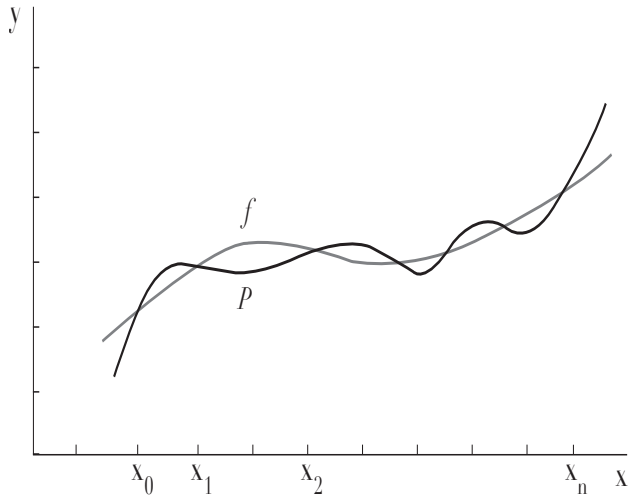


Figura 4.2: Función interpoladora de $n+1$ puntos

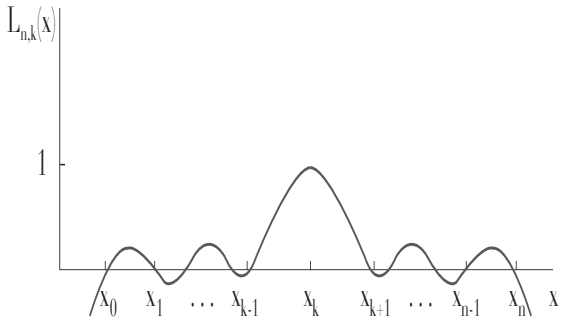


Figura 4.3: Polinomio de interpolación de grado n

$$L_{n,k}(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_{k-1})(x-x_{k+1})\dots(x-x_n)}{(x_k-x_0)(x_k-x_1)\dots(x_k-x_{k-1})(x_k-x_{k+1})\dots(x_k-x_n)}$$

para cada $k = 0, 1, \dots, n$.

Por tanto, dados $x_0, x_1, \dots, x_n$, $n+1$ puntos distintos y f una función desconocida pero de la que se conoce el valor que toma en esos puntos, puede decirse que $P_n(x)$ es el único polinomio de grado menor o igual que n que coincide con $f(x)$ en dichos puntos $x_0, x_1, \dots, x_n$. Esta fórmula no es adecuada desde el punto de vista numérico y de coste computacional, pues la modificación, eliminación o adición de un nodo hace inservibles a los polinomios de Lagrange ya calculados. Una estrategia

muy extendida es calcular el polinomio de interpolación "por etapas", más conocida como *Método de las diferencias divididas*. Una explicación en detalle de dicho método puede estudiarse en [42].

4.2.1.2. Interpolación de Hermite

Como se ha visto en la sección anterior, los polinomios de Lagrange coinciden con la función que se quiere aproximar en puntos prefijados. Por lo que serán una buena aproximación de la función que buscada cuando la información de partida sean valores exactos del riesgo en puntos determinados del atributo. Se plantea ahora una nueva situación, donde además de conocer cuanto vale el riesgo en un valor concreto del atributo, se conocen también ciertas propiedades que se cumplen en dicho punto o en un entorno del mismo. Estas propiedades pueden ser sobre el crecimiento o decrecimiento del riesgo en ese punto, la pendiente, etc. Por ejemplo, cuanto vale el riesgo en un punto y además si crece o decrece en un entorno del mismo. Luego lo se tendría información sobre cuanto vale la función en un punto y además cuál es el valor de su derivada en dicho punto.

Para poder calcular la función de interpolación que mejor se aproxime a esos datos, se propone el método interpolación de Hermite, que calcula el polinomio que coincide con la función buscada y su primera derivada en puntos prefijados. Este modelo es un poco más complicado que el de Lagrange, puesto que las condiciones adicionales de la derivada hacen que el grado del polinomio de Hermite sea mayor. Sean los valores de la función f, en $n+1$ puntos, $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))$, cuya derivada es continua en un intervalo $[a, b]$ que contiene los $n+1$ puntos de partida $x_0, x_1, \ldots, x_n$. El polinomio de Lagrange que coincide con f en estos puntos será en general de grado n. Si se requiere, además, que la derivada del polinomio de Hermite coincida con la derivada de f en $x_0, x_1, \ldots, x_n$, entonces es lógico pensar que las $n+1$ condiciones adicionales eleven el grado del polinomio de Hermite a $2n+1$. Esto hace que sea un poco más complicada su expresión puesto que ahora el espacio vectorial de los polinomios va a ser de grado no menor que $2n+1$. A continuación se verá formalmente como se define el polinomio de Hermite.

Definición 4.2. Sea $f \in C^1[a, b]$ y $x_0, x_1, \ldots, x_n$ $n+1$ puntos distintos del intervalo [a,b]. El único polinomio de menor grado posible que coincide con f y f' en $x_0, x_1, \ldots, x_n$ es el *polinomio de interpolación de Hermite* de grado menor o igual

que $2n+1$ dado por

$$H_{2n+1}(x) = \sum_{j=0}^{n} f(x_j) H_{n,j}(x) + \sum_{j=0}^{n} f'(x_j) \hat{H}_{n,j}(x),$$

donde

$$H_{n,j}(x) = [1 - 2(x - x_j) L'_{n,j}(x)] L^2_{n,j}(x)$$

y

$$\hat{H}_{n,j}(x) = (x - x_j) L^2_{n,j}(x)$$

Aquí, $L_{n,j}(x)$ denota el j-*ésimo* coeficiente polinomial de grado n de Lagrange.

Aunque la fórmula de Hermite proporciona una descripción completa de los polinomios de Hermite, la necesidad de determinar y evaluar los polinomios de Lagrange y sus derivadas hace que el procedimiento sea tedioso incluso para valores bajos de n. Por lo que para facilitar el cálculo se usa el *Método de las diferencias divididas*, igual que con el método anterior.

4.2.1.3. Interpolación con funciones Splines

Los métodos de interpolación presentados hasta ahora, aproximan la función buscada por un polinomio que verifica con exactitud los datos de partida. Por tanto, cuanta más información de el experto, mayor será tanto el número de puntos conocidos de esa función como el grado del polinomio, lo que se traduce en un modelo complejo que resulta poco natural, y por consiguiente en un aumento de los errores. Por estas razones y visto que calcular la función interpoladora es una buena solución para resolver el problema planteado, se propone un enfoque alternativo, que permita cometer menos errores. La idea es dividir el intervalo en un grupo de subintervalos y construir un polinomio de aproximación diferente sobre cada subintervalo, lo que se conoce como *aproximación polinomial a trozos.*

La aproximación polinomial a trozos más simple consiste en unir los puntos dados $(x_0, f(x_0)), (x_1, f(x_1)), \ldots, (x_n, f(x_n))$ mediante una serie de líneas rectas para formar una línea quebrada como la que se muestra en la Figura 4.4. Este modelo en general no es derivable en los extremos de los subintervalos, ver 4.4, la función que interpola no es "suave" en dichos puntos; sin embargo, a menudo ocurre

que las condiciones físicas del problema requieren suavidad, por lo que la función interpoladora debe ser derivable con continuidad. Esto hace que la aproximación polinomial a trozos más usada sea aquella que use polinomios de grado no muy alto en los subintervalos y además exista derivabilidad en los nodos. Esta es la base de la interpolación con *Splines*.

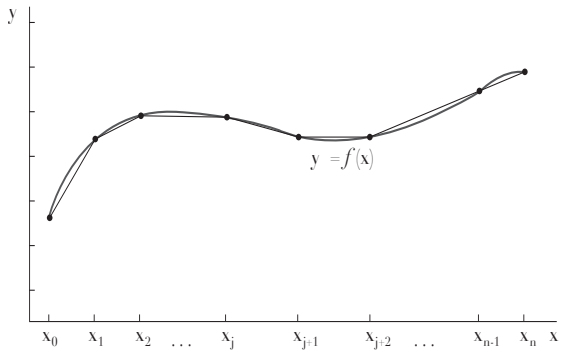


Figura 4.4: Aproximación polinomial a trozos mediante polinomios de grado 1

Definición 4.3. Sea $[a, b]$ un intervalo real, dados $a \leq x_0 < x_1 < \ldots < x_n \leq b$, $(n+1)$ puntos distintos de $\mathbb{R}$ y $f(x_0), f(x_1), \ldots, f(x_n)$, $n+1$ valores reales arbitrarios, se denomina función *spline de orden* p o *p-spline interpolador* de los datos $\{(x_i, f(x_i)), i = 0, \ldots, n\}$ a una función real S definida en el intervalo $[a, b]$ verificando:

a) $S \in C^{p-1}([a, b])$

b) En cada subintervalo $[x_i, x_{i+1}]$, S es un polinomio de grado p

c) $S(x_i) = f(x_i)$, $\forall i = 0, \ldots, n$

Las funciones splines de grado impar tienen propiedades de suavidad en sus gráficas, siendo las menos oscilantes, lo que las hace especialmente preferidas. Esto unido a que lo más simple es usar polinomios de grado bajo y a que las poligonales presentan ya discontinuidades en la derivada primera hace que las funciones splines más populares sean las cúbicas. Luego el spline más utilizado en la práctica es el spline de orden 3 o spline cúbico, una función de clase 2 donde cada trozo es un polinomio de grado 3. Para determinar de forma única al spline buscado se requieren dos condiciones adicionales; tres elecciones muy interesantes y también muy populares en la práctica son las siguientes:

- *Caso cúbico natural* $S''(x_1) = 0,\ S''(x_n) = 0$
- *Caso cúbico periódico* $S'(x_1) = S'(x_n),\ S''(x_1) = S''(x_n)$
- *Caso cúbico sujeto* $S'(x_1) = y'_1,\ S'(x_n) = y'_n$

4.2.1.4. ¿Cuándo es adecuado el uso de Interpolación?

Como se ha visto, los métodos de interpolación propuestos son una herramienta útil para resolver el problema planteado. Sin embargo, también resulta adecuado en esta situación poner de manifiesto que, aunque simples de analizar y fáciles de usar, lo métodos de interpolación, en particular Lagrange y Hermite, presentan problemas que obligan a actuar con precaución. Con el fin de facilitar la selección del método en función de las características del problema se analizará bajo que condiciones estos métodos son los más adecuados:

a) *Cantidad de información*

Lagrange y Hermite no siempre funcionan correctamente con cantidades mayores de seis puntos. A medida que crece el grado del polinomio interpolador y se usa alguno de estos dos métodos, se percibe una creciente variación entre puntos de información consecutivos, lo que produce que la aproximación entre dos puntos continuos sea muy distinta a la que se esperaría. Por tanto, aunque interpolar con los métodos de Lagrange y Hermite es una forma fácil de calcular una aproximación de la función buscada cuando se conocen valores puntuales y/o propiedades sobre dicha función, si el número de valores de partida es grande, el polinomio interpolador realmente no será una buena aproximación de la función. Una buena opción, en esta situación, son las funciones spline polinómicas de grado bajo, que garantizan un comportamiento satisfactorio de la función interpoladora. Además estas funciones admiten una mayor cantidad de información.

b) *Tipo de información*

El tipo de información que admiten estos métodos es numérica. De entre toda clase de información numérica de la que se puede disponer, estos métodos están especialmente indicados para ser utilizados con información que proceda principalmente de datos empíricos, como ejemplos, muestras, etc. La razón principal es que es que estos modelos tienen la característica de verificar con exactitud los valores de partida. Si además, el modelo elegido es Hermite, entonces también

admite información sobre propiedades de la función como puedan ser: crecimiento o decrecimiento, positividad, etc.

c) *Coste computacional y esfuerzo de implementación*

Su coste computacional es bajo y además, no requieren esfuerzo de implementación, puesto que vienen integrados en cualquier paquete matemático.

d) *Flexibilidad e interpretabilidad*

Se ajustan con exactitud a lo que el experto dice, por lo que no son muy flexibles. De entre todos, tal vez, las funciones spline sean las más flexibles con la información. Con respecto a la interpretabilidad, pese a ser sencillos de entender sobre todo si se ve su gráfica, requieren de ciertos conocimientos para interpretarlos por completo. Pese a esto, resulta fácil su uso.

e) *Propiedades matemáticas del modelo*

Con respecto al análisis matemático de los modelos de interpolación cabe resaltar las siguientes propiedades:

i) Existen siempre y son únicos

ii) No conservan ni el crecimiento, ni la concavidad, ni la negatividad, etc.

iii) Presentan fluctuaciones frente a pequeños cambios (Lagrange y Hermite) añadir o quitar un punto cambia mucho la aproximación, sin embargo spline es estable.

En resumen, son una herramienta útil y fácil de usar que se ajusta bien a los datos. La situación en que resulta más adecuado su uso es aquella en que la información provenga de ejemplos, si es muy poca usar Lagrange (o Hermite si esta información viene acompañada de ciertas propiedades), y si la cantidad de información es mayor entonces usar las funciones splines como función interpoladora.

4.2.2. Aproximación

La aproximación es otro método matemático que propone una solución para el problema de calcular el valor de una función en un punto cuando, o bien no se conoce la expresión de la función en ese punto, o dicha expresión no es fácil de evaluar en él. Consiste en sustituir esa función por otra más sencilla que sea "próxima" en un cierto sentido que se ha de precisar previamente, sin que necesariamente tenga que

coincidir en unos puntos dados como ocurre con el modelo de interpolación.

A rasgos generales el problema de aproximar una función f perteneciente a un conjunto de funciones F, consiste en encontrar elementos de un subconjunto U de F que sean "próximos" a la función f. Naturalmente, hay que precisar de alguna forma cómo se mide la mayor o menor proximidad de un elemento a otro, y de ahí que deba definirse en F una *métrica* o *pseudométrica*. Recordemos ahora, que una métrica en un conjunto X, no es más que una aplicación $d : X \times X \longrightarrow \mathbb{R}$ que verifica las siguientes condiciones:

a) $d(x, y) \geq 0 \ \forall x, y \in X$

b) $d(x, y) = d(y, x) \ \forall x, y \in X$

c) $d(x, y) \leq d(x, z) + d(z, y) \ \forall x, y, z \in X$

d) $d(x, y) = 0 \Rightarrow x = y$

Si se verifican a), b) y c) solamente, se tiene una pseudométrica. Tanto en un caso como en otro ya tiene sentido decir que un elemento v es más próximo a f que otro elemento w o a la inversa. Sin embargo, en este contexto, el caso que puede resultar más interesante es aquel en que F es un *espacio vectorial normado*. Esto quiere decir que F es un espacio vectorial sobre el cuerpo K real o complejo en el que hay definida una *norma*, es decir, una aplicación de F en $\mathbb{R}$ que suele denotarse $||\cdot||$ y que verifica las siguientes condiciones:

a) $||x|| \geq 0 \ \forall x \in F$

b) $||x + y|| \leq ||x|| + ||y|| \ \forall x, y \in F$

c) $||\lambda \cdot x|| = |\lambda| \cdot ||x|| \ \forall x \in F, \forall \lambda \in K$

d) $||x|| \Rightarrow x = 0$

Si falta la condición d) se dice que es una *seminorma*. En ambos casos la distancia entre dos elementos x e y viene expresada por $||x - y||$, y dará lugar a una métrica o pseudométrica según corresponda.

Teniendo en cuenta las características del problema planteado en este capítulo, la Aproximación es un método adecuado. Será indicado para resolver el problema

cuando la información disponible sean valores puntuales de la función buscada. A diferencia de la interpolación, este método no está obligado a tomar con exactitud dichos valores, sino que ha de hacerlo de forma tan próxima como se quiera, en el sentido de proximidad que se decida establecer, lo que se consigue imponiendo propiedades sobre el conjunto U. Veamos el problema de aproximación de forma abstracta:

Planteamiento general del Problema de Aproximación

Se entiende por *Problema General de Aproximación* (PGA) el que se plantea a continuación. Sea F un espacio vectorial normado sobre el cuerpo K real o complejo y U un subespacio vectorial de F que generalmente suele ser el conjunto de todos los polinomios. Resolver el PGA es buscar si existen elementos $u \in U$ tales que:

$$d(f, U) = inf_{v \in U} ||f - v||$$

Un elemento u de U que verifique esta condición, se dice que es una *mejor aproximación* de f en U.

Observar que no siempre existe mejor aproximación de f en U, puesto que el ínfimo, en general, no tiene por qué ser alcanzado para ningún elemento de U. Además, en caso de ser alcanzado ese ínfimo, puede ser el mismo para varios elementos u de U. Por tanto, la mejor aproximación de f en U no existe siempre y no tiene por qué ser única.

Asociada al tipo de información que el experto da, una buena forma de conocer el valor del riesgo, en ciertos puntos del dominio en los cuales no se conoce más que el valor del atributo, será la aproximación por mínimos cuadrados, en todas sus variantes. Veamos ahora de forma detallada cada uno de estos modelos.

4.2.2.1. Aproximación por Mínimos Cuadrados

Un problema interesante y muy estudiado dentro de la aproximación es el de la *Aproximación por Mínimos Cuadrados* en espacios vectoriales normados. De forma que fijado un espacio normado E, buscar la mejor aproximación por mínimos cuadrados de un elemento f en un subconjunto de ese espacio, equivale a hallar un elemento de dicho subconjunto que minimice la siguiente expresión:

$$||f - v||^2 =< f - v, f - v > \tag{4.2}$$

es decir, hallar un elemento que minimice el cuadrado de $f - v$ con el producto escalar del espacio E.

Por tanto, se llamará problema de *aproximación por mínimos cuadrados* a todo problema de aproximación en espacios normados. Sin embargo, se conocen tradicionalmente con ese nombre los problemas correspondientes a los casos particulares relacionados con los productos escalares definidos en Ecuación 4.3 y Ecuación 4.4. Dichos productos dan lugar respectivamente a la aproximación por mínimos cuadrados continua y a la aproximación por mínimos cuadrados discreta. Ambas se estudiarán, en detalle, a continuación.

$$< f, g >= \int_a^b f(x)g(x)dx \tag{4.3}$$

$$< f, g >= \sum_{i=1}^{n} f(x_i)g(x_i) \tag{4.4}$$

1. Aproximación por Mínimos Cuadrados Continua

Consideremos el espacio vectorial de las funciones reales continuas en un intervalo $[a, b]$ cerrado y acotado, es decir, en un conjunto compacto, con el producto escalar definido en la Ecuación 4.3 que hace de él un espacio prehilbertiano que se llamará E. En estas condiciones el teorema de existencia y unicidad de aproximación en espacios normados, asegura que para toda función f continua en $[a, b]$ existe una única mejor aproximación u en un subespacio $H \subseteq E$ de dimensión finita n, caracterizada en este caso, por verificar la siguiente condición:

$$\int_a^b (f(x) - u(x))v(x)dx = 0 \quad \forall v \in H \tag{4.5}$$

Al ser la función u mejor aproximación verifica Ecuación 4.6 y por tanto se dice que u es la *mejor aproximación por mínimos cuadrados continua* de la función f en H.

$$\int_a^b (f(x) - u(x))^2 dx = min_{v \in H} \int_a^b (f(x) - v(x))^2 dx \tag{4.6}$$

2. Aproximación por Mínimos Cuadrados Continua Ponderada

Un caso particular de la aproximación por mínimos cuadrados continua es la aproximación por mínimos cuadrados continua ponderada. Es un modelo bastante útil y realista, puesto que resulta intuitivo pensar que no todos los valores sobre los que de información el experto tendrán la misma importancia, y podría en algunos casos ser conveniente darles mayor peso. A continuación se estudiará este modelo, para ello se define el siguiente producto escalar:

$$< f, g >= \int_I w(x)f(x)g(x)dx \tag{4.7}$$

donde I es un intervalo cualquiera, $w(x)$ una función fija, llamada *función peso*, tal que $w(x) \geq 0$ en I, pudiendo ser nula sólo en un número finitos de puntos de I y tal que

$$\int_I w(x)f(x)dx \tag{4.8}$$

exista para toda función continua f.

Consideremos el espacio vectorial de las funciones reales continuas en un intervalo I con el producto escalar 4.7 que hace de él un espacio prehilbertiano que se llamará E. Igual que ocurría en el caso anterior, en estas condiciones el teorema de existencia y unicidad de aproximación en espacios normados, asegura que para toda función f continua en I existe una única mejor aproximación u de f en un subespacio $H \subseteq E$, caraterizada por:

$$\int_I w(x)(f(x) - u(x))v(x)dx = 0 \quad \forall v \in H \tag{4.9}$$

Al ser la función u mejor aproximación verifica Ecuación 4.10 y por tanto se dice que u es la *mejor aproximación por mínimos cuadrados continua ponderada* de la función f en H.

$$\int_I w(x)(f(x) - u(x))^2dx = min_{v \in H} \int_I w(x)(f(x) - v(x))^2dx \tag{4.10}$$

Este problema puede interpretarse diciendo que a cada punto x del intervalo $[a, b]$ se le conoce un "peso" $w(x)$ a la hora de calcular la norma de $f(x)$, con objeto de que los valores de esta función en unos puntos influyan más o menos que otros. El caso más sencillo corresponde a $I = [a, b]$, con $w(x) \equiv 1$. Este sería el del apartado anterior.

3. Aproximación por Mínimos Cuadrados Discreta

Consideremos el espacio vectorial $E = \mathbb{R}^n$ que con el producto escalar 4.4, forman un espacio vectorial normado. Dados $v^{(1)}, v^{(2)}, \ldots, v^{(n)}$ n vectores linealmente independientes de $\mathbb{R}^m$ y H el subespacio que engendran. Se dice que

$$u = \sum_{i=1}^{n} \lambda_i v^{(i)} \tag{4.11}$$

es la *mejor aproximación por mínimos cuadrados discreta* de f en H si

$$||f - u|| = min_{v \in H} ||f - v|| \tag{4.12}$$

es decir si

$$\sum_{k=1}^{m} (f_k - u_k)^2 = min_{v \in H} \sum_{k=1}^{m} (f_k - v_k)^2. \tag{4.13}$$

Este problema es completamente análogo al continuo, y un nuevo caso particular de la aproximación en espacios prehilbertianos. Por tanto, una vez más, u viene caracterizada por la condición de ortogonalidad de $f - u$.

4. *Aproximación por Mínimos Cuadrados Discreta Ponderada*

Como caso particular de la aproximación por mínimos cuadrados discreta, se tiene la aproximación por mínimos cuadros discreta ponderada, que al igual que en el caso continuo, será útil cuando el experto informe de antemano cuáles son los valores que considera más importantes. El modelo es igual al caso discreto pero con el producto escalar siguiente:

$$< f, g >= \sum_{k=1}^{m} w(x_k) f(x_k) g(x_k) \tag{4.14}$$

donde $w(x_k) > 0$ para todo k y se llama *función peso.*

4.2.2.2. Aproximación mediante Curvas Bézier

La aproximación mediante curvas Bézier, es otro modelo del que está muy extendido su uso. Uno de los inconvenientes de la interpolación polinómica era que si los datos de partida eran por ejemplo, convexos o positivos, el interpolante no siempre conservaba este tipo de propiedades. Desde el punto de vista de las aplicaciones, este

hecho motiva la búsqueda de una representación más adecuada, es decir, una base de polinomios bien adaptada. Ésta es la formada por los polinomios de Bernstein.

Definición 4.4. Sean $n, r \in \mathbb{N}$ tales que $0 \leq r \leq n$, se define el r-ésimo polinomio de Bernstein de orden n como:

$$B_r^n(t) = \binom{n}{r}(1-t)^{n-r}t^r$$

recorriendo t el intervalo $[0,1]$.

Los polinomios de Bernstein surgen al desarrollar por la fórmula del binomio de Newton la expresión $((1-t)+t)^n$, y presentan ciertas propiedades algunas de las cuales resultan interesantes:

1. El conjunto $\{B_r^n, 0 \leq r \leq n\}$ es una base del espacio de los polinomios de grado menor o igual que n, $\mathbb{P}_n$

2. $B_r^n(0) = B_r^n(1) = 0,\ r \neq 0,\ r \neq n$

   $B_0^n(0) = B_n^n(1) = 1$

   $B_0^n(1) = B_n^n(0) = 0$

3. $B_r^n(t) \geq 0\ \ \forall t \in [0,1]$

4. Los polinomios de Bernstein de grado n constituyen una partición de la unidad, es decir, para todo $t \in [0,1]$ se cumple que:

$$\sum_{r=0}^{n} B_r^n(t) = 1$$

5. $B_i^n(t) = (1-t)B_i^{n-1}(t) + tB_{i-1}^{n-1}(t)$

6. $max_{t\in[0,1]}B_r^n(t) = B_r^n(\frac{r}{n})$

7. $B_r^n(t) = B_{n-r}^r(1-t)$

Con la base de Bernstein se pueden construir curvas polinómicas, que denominaremos *Curvas de Bézier*.

Definición 4.5. Sean $y_i, 0 \leq i \leq n$ números reales, la función polinómica $f(t) = \sum_{i=0}^{n} y_i B_i^n(t), t \in [0,1]$, tiene como gráfica una curva que puede parametrizarse como

$$X(t) = (t, f(t)) = \sum_{i=0}^{n} \binom{\frac{i}{n}}{y_i} B_i^n(t) = \sum_{i=0}^{n} b_i B_i^n(t),$$

sabiendo $b_i = \binom{\frac{i}{n}}{y_i}$.
si, en general, los puntos b_i son puntos arbitrarios de $\mathbb{R}^k$, $k = 2, 3$, las funciones

$$X(t) = \sum_{i=0}^{n} b_i B_i^n(t)$$

son curvas parametrizadas en $\mathbb{R}^k$ con funciones coordenadas polinómicas. En los tres casos indicados los puntos b_i se denominan puntos de Bézier o de control, y la línea poligonal que determinan se llama polígono de Bézier, de control o B-polígono.

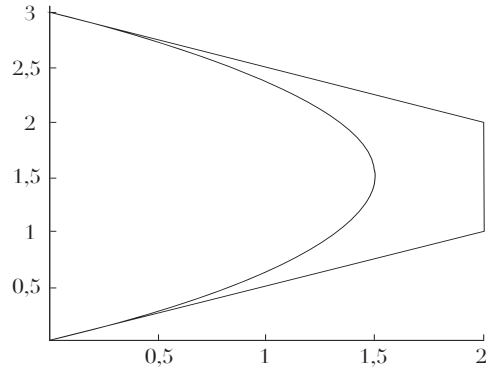


Figura 4.5: Curva de Bézier y su polígono de control. Los puntos de control son $b_0 = (0,0)$, $b_1 = (2,1)$, $b_2 = (2,2)$, $b_3 = (0,3)$

Es sencillo comprobar que la curva de Bézier así definida interpola los puntos b_0 y b_n y es tangente a los segmentos inicial y final del polígono de control, respectivamente. Las curvas Bézier presentan una propiedad geométrica muy importante, y es que sus trazas están contenidas en las envolventes convexas de sus puntos de control, lo que se deduce de la no negatividad de los polinomios de Bernstein. Otra propiedad interesante de estas curvas es que presentan la propiedad de disminución de la variación, que establece que el número de puntos de corte de una línea recta

con una curva de Bézier plana es a lo sumo igual al número de puntos de intersección con su polígono de control. La mayor ventaja en nuestro contexto es que el polígono de control de una curva Bézier sugiera la forma de ésta ya que hace posible que se pueda modificar interactivamente la gráfica moviendo alguno de los puntos de control. De hecho, si un punto de Bézier, b_i, se mueve a una posición, $\overline{b_i}$, entonces todos los puntos de la curva se mueven hacia $\overline{b_i}$ en una dirección paralela a $\overline{b_i} - b_i$. Así pues los cambios en la gráfica pueden ser previstos. Lo cual nos ayuda a rectificar cualquier información en cualquier momento con poco trabajo. Es una ventaja importante, teniendo en cuenta que la información sobre la que se trabaja es subjetiva.

4.2.2.3. ¿Cuándo es adecuado el uso de Aproximación?

Al igual que los modelos de interpolación, los de aproximación también resultan una herramienta útil para resolver el problema planteado. Su desventaja principal es que su existencia no está siempre asegurada, lo que obliga a poner ciertas restricciones al modelo y por tanto pedir información adicional al experto. A continuación con el fin de facilitar la selección del método en función de las características que tiene el problema se analizará bajo que condiciones estos métodos son los más adecuados:

a) *Cantidad de información*

   En general la aproximación funciona bien con mucha información, para estos modelos cuanto más información de partida se tenga mejor será su resultado.

b) *Tipo de información*

   El tipo de información que admiten estos métodos es numérica. De entre toda clase de información numérica de la que se pueda disponer, estos métodos están especialmente indicados para ser utilizados con información que proceda directamente del experto, o que por las razones que sean pudiera ser no exacta; debido a que este modelo no verifica la información con exactitud, sino que se aproxima tanto como se quiera y según el criterio que previamente se establezca. Estos modelos no admiten información en forma de propiedades de la función como ocurría con la interpolación.

c) *Coste computacional y esfuerzo de implementación*

   Su coste computacional es bajo y además, no requieren esfuerzo de implementación, puesto que vienen integrados en cualquier paquete matemático.

d) *Flexibilidad e interpretabilidad*

Se ajustan a lo que el experto dice en el sentido de proximidad que sea fijado previamente. En este sentido, la aproximación resulta más flexible que la interpolación. Con respecto a la interpretabilidad, pese a ser sencillos de entender sobre todo si se ve su gráfica, requieren de ciertos conocimientos para interpretarlos por completo. Pese a esto, resulta fácil su uso.

e) *Propiedades matemáticas del modelo*

Con respecto al análisis matemático de los modelos de interpolación cabe resaltar las siguientes propiedades:

I) No siempre está garantizada su existencia, y para asegurarla hay que exigir ciertas restricciones, lo que limita el modelo.

II) Si existe, no tiene por qué ser única. Esto realmente no presenta una desventaja, pues se busca una función que aproxime bien los datos y no importa que exista más de una.

III) En relación a los problemas que presentaba la interpolación, con respecto a la herencia de convexidad, positividad, etc. de la función de partida; las curvas Bézier juegan un papel muy destacado, ya que gracias a las propiedades que presentan los polinomios de Bernstein, ellas sí heredan este tipo de propiedades.

IV) Es estable frente a pequeños cambios puntuales, lo cual supone una ventaja frente a la interpolación. Esto proporciona facilidad a la hora de expresarse el experto o de rectificar si fuera necesario.

En resumen, son una herramienta útil y fácil de usar que se ajusta bien a los datos. La situación en que resulta más adecuado su uso es aquella en que la información es dada directamente por el experto y no procede de muestras o datos. Cuanta más información se tenga mejor será la aproximación del modelo.

4.2.3. Modelado difuso de la información

El modelado difuso de datos está basado en la lógica difusa [194] y también da solución al problema de cómo aproximar una función a partir de cierta información dada. Construir un modelo difuso de datos, es definir un número razonable de reglas difusas del tipo *si-entonces* ("if-then"), capaz de aproximar la función o relación funcional buscada a partir de la información disponible, con la gran ventaja de que

dicha información puede ser tanto numérica como lingüística. Es otra herramienta interesante con la que resolver el problema planteado, principalmente cuando no solo se dispone de información lingüística. Actualmente, las reglas difusas son muy recomendadas como herramienta para expresar o representar conocimiento. En esta memoria, se ha seleccionado Wang y Mendel [**?**] como algoritmo para generar reglas difusas por ser el método más simple y a la vez más extendido en la práctica. Veamos en detalle su funcionamiento.

4.2.3.1. Wang-Mendel: Generando reglas difusas a partir de ejemplos

Wang y Mendel es un método cuyo uso está muy extendido y con el que pese a su sencillez se han obtenido buenos resultados. Tiene la enorme ventaja de que partiendo de información numérica y lingüística puede mezclarlas y aproximar cualquier función real sobre un conjunto compacto con bastante precisión. A continuación se desarrollan en detalle los cinco pasos que forman el procedimiento para generar la base de reglas difusas a partir de los datos. Se utilizará un conjunto pequeño de datos con dos entradas y una sola salida, para así simplificar su entendimiento. Es inmediata su extensión a una dimensión mayor.

Sean $\{(x_1^{(1)}, x_2^{(1)}; y^{(1)}), (x_1^{(2)}, x_2^{(2)}; y^{(2)}), \ldots\}$ el conjunto de parejas de datos, dado por el experto, donde x_1 y x_2 son las entradas e y la salida del sistema. Se generará un conjunto de reglas difusas con los datos entrada-salida fijados para aproximar, $f : (x_1, x_2) \longrightarrow y$, que es la función o relación funcional buscada. Veamos a continuación los pasos a seguir.

Paso 1. Dividir los espacios de Entrada y Salida en regiones difusas.

Supongamos que x_1, x_2 e y toman valores con una alta probabilidad en los intervalos $[x_1^-, x_1^+]$, $[x_2^-, x_2^+]$ y $[y^-, y^+]$ respectivamente, aunque también es posible que los tomen fuera. Se divide cada intervalo en $2N + 1$ regiones y se asigna a cada región una función de pertenencia, $m(x_1)$, $m(x_2)$ y $m(y)$. La Figura 4.6 muestra un ejemplo gráfico de como podrían ser estas particiones y puede verse como x_1 se divide en cinco regiones ($N = 2$), x_2 en siete ($N = 3$) e y en cinco ($N = 2$). La forma de las funciones de pertenencia es triangular. No olvidemos que se trata de un ejemplo y que se puede utilizar cualquier división de los intervalos y cualquier función de pertenencia.

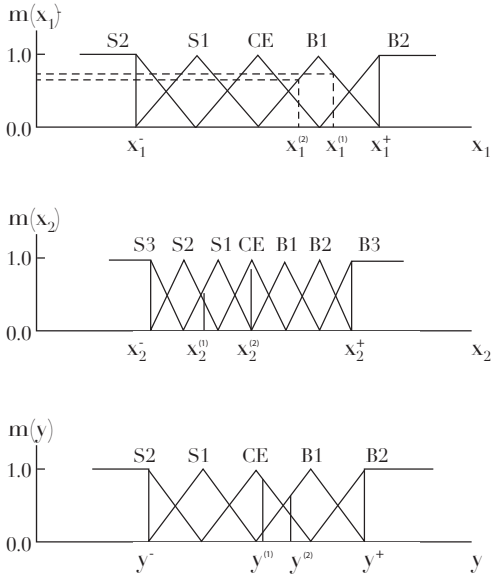


Figura 4.6: Ejemplo de una división de los espacios de entrada y salida en regiones difusas y sus correspondientes funciones de pertenencia

Paso 2. Generar reglas difusas a partir de las parejas de datos.

Las reglas difusas se generan en dos pasos y haciendo uso de la función de pertenencia. Primero, se calculan los grados de pertenencia de $x_1^{(i)}$, $x_2^{(i)}$ e $y^{(i)}$ en las diferentes regiones, como puede verse en la Figura 4.6 $x_1^{(1)}$ tiene grado de pertenencia 0.8 en *B1*, 0.2 en *B2* y cero en las otras regiones.

Segundo, se asigna a cada $x_1^{(i)}$, $x_2^{(i)}$ e $y^{(i)}$, la región cuyo grado de pertenencia sea máximo. Por ejemplo, observando la Figura 4.6 a $x_1^{(1)}$ se le asigna *B1* puesto que en esa región es donde toma el valor máximo. Repitiendo este procedimiento se obtiene una regla por cada par de valores entrada-salida:

- $(x_1^{(1)}, x_2^{(1)}; y^{(1)}) \Rightarrow [x_1^{(1)}$(valor max es 0.8 en *B1*), $x_2^{(1)}$ (valor max es 0.7 en *S1*); $y^{(1)}$(valor max es 0.9 en *CE*)]. Por tanto, *Regla 1* quedaría:

  Regla 1: Si x_1 es *B1* y x_2 es *S1*, entonces y es *CE*

- $(x_1^{(2)}, x_2^{(2)}; y^{(2)}) \Rightarrow [x_1^{(2)}$(valor max es 0.6 en *B1*), $x_2^{(2)}$ (valor max es 0.7 en *CE*); $y^{(2)}$(valor max es 0.9 en *B1*)]. Por tanto, *Regla 2* queda:

Regla 2: Si x_1 es *B1* y x_2 es *CE*, entonces y es *B1*

Las reglas generadas según este procedimiento usan el operador "y" (en inglés *and*).

Paso 3. Asignar un grado a cada regla para eliminar reglas conflictivas.

Al tener distintos pares de datos y como cada par genera una regla, es muy probable que existan conflictos entre ellas. Se dirá que dos reglas son conflictivas si tienen el mismo antecedente y distinto consecuente. Para eliminar las reglas conflictivas se calcula el grado de cada regla y se selecciona de entre todas aquella que tenga grado máximo. Existen varios métodos para calcular el grado de una regla, de entre ellos se han seleccionado:

$$D_1(Regla) = m_A(x_1) \cdot m_B(x_2) \cdot m_C(y) \tag{4.15}$$

$$D_2(Regla) = m_A(x_1) \cdot m_B(x_2) \cdot m_C(y) \cdot m^{(1)} \tag{4.16}$$

dónde $m^{(1)}$ es la creencia que tenemos nosotros sobre su *utilidad* y *A, B, C* las regiones de la función de pertenencia correspondientes. En los experimentos se utilizará el método definido en la Ecuación 4.15. Sin embargo, cuando en la práctica se cuente con información a priori, se aplicará el método definido por la Ecuación 4.16. En esta situación el experto dirá cuales son los pares de valores más útiles. Esto, en aplicaciones reales, puede ser muy importante, debido a que no todos los datos recogidos serán igual de informativos. Ambas estrategias son valiosas.

Paso 4. Crear una base de reglas difusas combinadas.

La forma que presenta generalmente una base de reglas difusas es la que puede verse en la Figura 4.7, para rellenarla se sigue la siguiente estrategia: se unen a la base de reglas difusas tanto las reglas generadas por los datos numéricos como las reglas generadas por la información lingüística dada por un experto. Si hubiera más de una regla en una casilla, gana la regla de grado máximo. De este modo la información numérica y lingüística queda codificada dentro de una misma estructura. Si

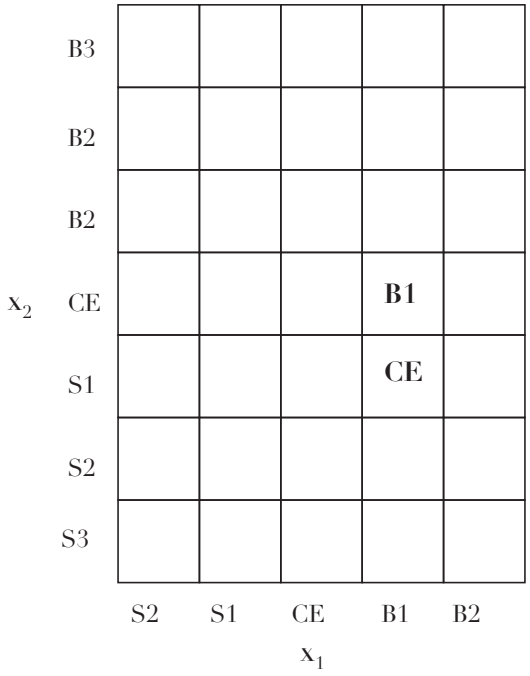


Figura 4.7: Base de reglas difusas

el operador de las reglas es el operador y (como ocurre en este caso) por cada regla solo se rellena una casilla. Véase un ejemplo en Figura 4.7.

Paso 5. Calcular la correspondencia basada en la base de reglas difusas combinadas.

Para ello se "desfuzzifica" el resultado, como estrategia para tal fin se usa la siguiente: primero, dada la entrada (x_1, x_2), se combinan los antecedentes de la i-ésima regla usando el producto para calcular el grado de cada regla según la Ecuación 4.15

$$m^i_{O^i} = m_{I^i_1}(x_1) \cdot m_{I^i_2}(x_2) \tag{4.17}$$

donde O^i es la región de salida de la Regla i e I^j_i denota la región de entrada de la Regla i para la componente j-ésima. Después se usa la fórmula de desfuzzificación del centroide, Ecuación 4.18, para calcular el grado final de la salida:

$$y = \frac{\sum_{i=1}^{k} m^i_{O^i} \overline{y}^i}{\sum_{i=1}^{k} m^i_{O^i}} \tag{4.18}$$

donde $\overline{y}^i$ denota el valor central de la región O^i y k el número de reglas en la base de reglas.

4.2.3.2. ¿Cuándo es adecuado el modelado difuso de los datos?

Wang y Mendel es un método general cuyo uso está muy extendido y que ha sido utilizado en multitud de aplicaciones con éxito. Genera reglas difusas a partir de datos numéricos y/o lingüísticos, con la enorme ventaja de ser capaz de mezclar estos dos tipos de información y aproximar cualquier función real sobre un conjunto compacto con bastante precisión. Este método resulta muy interesante y además, es muy útil para el experto a la hora de expresar su conocimiento. Igual que con los otros modelos presentados en este capítulo, se estudiaran los aspectos del problema que ayuden a decidir cuando será más adecuado utilizarlo.

a) *Cantidad de información*

   Este método admite tanta información como se quiera, y cuánta más información mejor será el resultado.

b) *Tipo de información*

   Este método admite información numérica y lingüística. Esta es su principal ventaja, ya que en la mayoría de problemas en los que la información viene del exterior esta puede obtenerse de dos formas distintas: de ejemplos, luego será información de tipo numérica, o del conocimiento del experto, para el que la forma más natural de expresarse será en forma lingüística. En bastantes ocasiones, estos dos tipos de información por separado pueden ser insuficientes, ya que aunque el sistema sea controlado por un experto, alguna información puede perderse al expresar su conocimiento de forma lingüística. Del mismo modo, la información que sólo sea extraída de muestras o ejemplos puede ser insuficiente también, pues resulta difícil que recoja todos los casos. Por tanto, entre todos los modelos presentados es el que más información del problema admite.

c) *Coste computacional y esfuerzo de implementación*

   Su coste computacional es bajo y su implementación no requiere demasiado esfuerzo.

d) *Flexibilidad e interpretabilidad*

De entre todos los métodos este es el que mayor grado de flexibilidad e interpretabilidad presenta. Los otros modelos, frente a este, son poco flexibles, principalmente por la rigurosidad matemática que implican, lo cual dificulta tanto la forma en que debe pedirse la información al experto, como la forma en que éste pueda expresar su conocimiento y también, en como se interprete el resultado que devuelve el modelo. Además, es muy flexible a la hora de su construcción, ya que existe mucha libertad al elegir las funciones de pertenencia, admite cualquier tipo de información y permite tener en cuenta otro tipo de detalles. Esta flexibilidad se debe en gran parte a la lógica difusa, ya que es esa característica una de las aportaciones más importantes de ésta al modelado de sistemas, permitiendo una descripción del conocimiento entendible para el ser humano, facilitando la forma en que se expresa el experto y haciendo más interpretables las relaciones entre la entrada y la salida, gracias al uso de reglas lingüísticas similares a las que podría emplear un experto humano. La interpretabilidad por tanto es otra de las grandes ventajas de este modelo.

e) *Propiedades Matemáticas del Modelo*

Con respecto al análisis matemático, cabe resaltar las siguientes propiedades:

i) Existe siempre.

ii) Es estable frente a pequeños cambios puntuales igual que ocurría con la aproximación. Esto supone una ventaja frente a la interpolación. Además proporciona facilidad a la hora de expresarse el experto o de rectificar si fuera necesario.

En resumen, es un modelo fácil de usar y que facilita el trabajo tanto al experto como al usuario final. Sus principales características son las flexibilidad y la interpretabilidad, heredadas principalmente de la lógica difusa. Podría usarse en cualquier ocasión, pero resulta especialmente conveniente cuando tengamos información numérica y lingüística del problema o simplemente información lingüística.

4.3. Asignación de riesgo en ausencia total del experto

Como ya se ha comentado, contar constantemente con un experto es muy complicado, de hecho simplemente conseguir información específica, ya es una tarea difícil.

Por lo tanto, y para poder usar la función IRL con independencia total del experto, en esta sección se presenta un nuevo concepto llamado *Riesgo Local Probabilístico* [26]. Este concepto es similar al de Información de Riesgo, tal vez no tan preciso, pero con la ventaja de ser asignado de forma automática. Para ello, en lugar de utilizar la opinión del experto o información proveniente de fuentes externas como muestras o ejemplos, se hace uso de la información que dan los casos almacenados en la base de casos.

4.3.1. Un método automático para estimar la función Información de Riesgo Local

Cuando no se dispone de ninguna información de partida y aún así se está interesado en utilizar la IRL, se define un nuevo concepto llamado Riesgo Local Probabilístico [26]. Éste, al igual que la IRL, da información acerca de lo apropiada que es una solución para resolver un problema, pero con la ventaja de que es asignado de forma automática, en lugar de coger la información del exterior, hace uso de los casos almacenados en memoria. A continuación se define formalmente este concepto.

Definición 4.6. Sea S una de las soluciones almacenadas en memoria y $A_i = a_i$ uno de los atributos del caso que toma el valor concreto a_i. Se define el riesgo local probabilístico para el i-ésimo atributo, $PR_i(A_i)$, como:

$$PR_i(A_i) = \frac{\text{Número de casos con solución } S \text{ cuyo atributo } A_i \text{ verifica P}}{\text{Número total de casos en la base de casos cuya solución asociada es } S}$$

donde P=$\{A_i = \nu \ / \ |\nu - a_i| \leq \alpha, \alpha \in \mathbb{R}\}$, es decir, el conjunto de atributos que dista de a_i menos de la cantidad fijada α.

Aunque a primera vista el riesgo local probabilístico es menos preciso que la opinión del experto (puesto que éste solo toma información de los casos almacenados en memoria) también presenta ventajas:

1. Es una verdadera función de probabilidad
2. Si el caso recuperado es finalmente almacenado en la base de casos, PR_i aumentará, debido a que la próxima vez que sea calculado habrá un caso más en la base de casos cuyo atributo A_i verificará la propiedad P. En este sentido se podrá hablar de aprendizaje.

3. Se asigna de forma automática, lo que simplifica bastante el uso del riesgo en cualquier problema.

En este caso, el valor se obtiene de forma numérica, lo que no supone ninguna restricción respecto a la asignación lingüística, puesto que podemos seguir el mismo proceso que con la similitud y los pesos que son asignados numéricamente.

4.3.1.1. ¿Cuándo es adecuado utilizar el riesgo probabilístico?

Este modelo está especialmente diseñado para el caso en que no contemos con ninguna información extra del problema. Aun así, para que su estudio sea paralelo al de los otros métodos, se estudiarán los aspectos a destacar del problema.

a) *Cantidad de información*

Está diseñado para ser usado sin información de partida, la información que usa es la que el propio método extrae de los casos almacenados en la base de casos, por lo que cuanto más casos haya almacenados mejor será el resultado obtenido.

b) *Tipo de información*

Como no necesita de información de partida no es determinante el tipo.

c) *Coste computacional y esfuerzo de implementación*

Su coste computacional es bajo, y además su implementación no requiere de mucho esfuerzo.

d) *Flexibilidad e interpretabilidad*

Es tan flexible como los casos almacenados en la base de casos le permitan serlo, y además tiene una fácil interpretación, no se requiere grandes conocimientos para ello.

e) *Propiedades matemáticas del modelo*

Con respecto a las propiedades matemáticas de este modelo, se puede decir que, como ya se ha comentado:

I) Es una verdadera función de probabilidad.

II) Permite hablar de aprendizaje, ya que si el caso recuperado es finalmente almacenado en la base de casos, PR_i aumentará, debido a que la próxima vez que sea calculado habrá un caso más en la base de casos cuyo atributo A_i verifique la propiedad P.

4.4. Resultados experimentales: Estudio comparativo

4.4.1. Descripción del problema: Casos de estudio

En esta sección se prueban empíricamente los modelos presentados en el capítulo, para así conocer su comportamiento y extraer conclusiones sobre las situaciones en que será más adecuado su uso. Con cada método se realizó un estudio comparativo en distintas etapas y bajo diferentes hipótesis. Las hipótesis simulan posibles escenarios y las etapas marcan la cantidad de información disponible. No se han utilizado todos los modelos sino sólo aquellos que se ajustaban al problema concreto según las premisas supuestas. Todos los resultados se compararon con los obtenidos en el Capítulo 3 para comprobar en qué grado las aproximaciones se ajustan a la realidad.

4.4.1.1. Los datos

Se han utilizado las mismas bases de datos del capítulo anterior para estudiar empíricamente el comportamiento de los modelos de asignación de riesgo propuestos. Es más, en cada prueba se usaron las mismas particiones y los mismos conjuntos de prueba de dicho capítulo. Una descripción detallada de estos puede encontrarse en la Sección 3.4.1.1 del Capítulo 3.

4.4.1.2. Ejecución de los experimentos

Las pruebas de análisis y comportamiento de los modelos se caracterizarán por las etapas e hipótesis establecidas. Mientras que las etapas quedarán marcadas por la cantidad de información, las hipótesis, simularán distintos escenarios. Para cada situación se seleccionará el representante que se considere más adecuado. Se decidió esta forma de proceder porque permitirá obtener conclusiones reales acerca del comportamiento de los modelos y optimizar los resultados. No tendría sentido, por ejemplo, utilizar para una prueba interpolación si la hipótesis indica que la información de partida es lingüística; tampoco sería adecuado utilizar interpolación de Lagrange si la hipótesis dice que se dispone de mucha información, puesto que en este caso dicho modelo no es estable, etc. Utilizar el modelo adecuado ayudará a ver en que grado las aproximaciones propuestas se ajustan a la realidad, lo que permitirá obtener apoyo experimental a los criterios de selección de métodos propuestos. Las hipótesis establecidas son:

- *Hipótesis 1* : La información es puntual y proviene de muestras reales o ejemplos y/o además el experto aconseja que debe verificarse con exactitud.
- *Hipótesis 2*: La información es imprecisa, procede de alguna fuente no muy fiable y/o el experto aconseja que es suficiente con que se aproxime a los datos tanto como se considere conveniente. Permite ser flexibles con respecto a la información.
- *Hipótesis 3*: La información está en forma lingüística y numérica o simplemente lingüística.
- *Hipótesis 4* : No se dispone de ninguna información.

Para cada posible escenario se selecciona el modelo más adecuado, con el objetivo de que la simulación sea lo más real posible. Bajo la *Hipótesis 1*, se utiliza interpolación ya que por sus propiedades es el método que verifica la información de partida con total exactitud. De entre todos los métodos de interpolación se selecciona el de funciones spline por se el más estable cuando la cantidad de información crece. Bajo la *Hipótesis 2*, puede utilizarse tanto aproximación como Wang y Mendel. Entre todos los posibles métodos de aproximación propuestos, se selecciona mínimos cuadrados discreta (por el tipo de datos). Bajo la *Hipótesis 3* se utiliza Wang y Mendel porque es el único método capaz de partir de información lingüística. Finalmente, bajo la *Hipótesis 4* se utiliza el riesgo probabilístico, ya que ha sido especialmente diseñado para cubrir esta situación. Las etapas que marcarán nuestros experimentos serán las siguientes:

- *Etapa 1* : Se dispone del 15 % de la información total
- *Etapa 2*: Se dispone del 30 % de la información total
- *Etapa 3*: Se dispone del 50 % de la información total
- *Etapa 4* : Se dispone del 100 % de la información total

Para garantizar la fiabilidad de los experimentos se tomó de forma aleatoria con MATLAB 7.1 el 15 %, el 30 % y el 50 % de los casos de la base total. Esto se hizo de forma proporcional en función de las clases en que está dividida cada base. De estos casos se tomaron los valores de riesgo asociados y actuaron como ejemplos de partida para generar las funciones de riesgo, utilizando los métodos descritos en este capítulo. La Tabla 4.1 muestra el tamaño de los conjuntos de ejemplos de los que se partió

en cada etapa y por cada base. Se ha fijado como valor máximo 50 % porque más de esa cantidad es casi obtener toda la información del experto, debido a que los atributos no tienen un amplio intervalo de valores, sino que suelen tomar valores repetidos.

Tabla 4.1: Proporción de los conjuntos de ejemplos en cada etapa

	BUPA Liver	**German credit**	**Wine Recong**	**Glass Indetif**
Etapa 1	Total:51 (C1:21,C2:30)	Total:150 (C1:105,C2:45)	Total:32 (C1:24,C2:8)	Total:26 (C1:9,C2:10,C3:7)
Etapa 2	Total:103 (C1:43,C2:60)	Total:300 (C1:210,C2:90)	Total:64 (C1:49,C2:15)	Total:54 (C1:18,C2:22,C3:14)
Etapa 3	Total:175 (C1:75,C2:100)	Total:500 (C1:350,C2:150)	Total:107 (C1:81,C2:26)	Total:89 (C1:29,C2:36,C3:24)
Etapa 4	Total:345 (C1:145,C2:200)	Total:1000 (C1:700,C2:300)	Total:214 (C1:163,C2:31)	Total:178 (C1:59,C2:71,C3:48)

En cada etapa y para cada modelo se usó validación cruzada de 10 subconjuntos. Se realizaron 10 validaciones completas y se evaluaron los resultados según la precisión obtenida por cada modelo en cada prueba. Se utilizó la misma metodología descrita en la Sección 3.4.1.2 del Capítulo 3. De los modelos de Interpolación y Aproximación se utilizó la implementación que de ellos existe en MATLAB 7.1 mientras que Wang y Mendel y el Riesgo Probabilístico fueron implementados también en MATLAB 7.1.

4.4.2. Resultados finales del estudio

En esta subsección, se presentan los resultados obtenidos en cada etapa y bajo las distintas hipótesis establecidas. La precisión total obtenida bajo la primera hipótesis puede verse en la Tabla 4.2, la obtenida bajo la segunda en Tabla 4.3 y la obtenida bajo la 3 y la 4 en la Tabla 4.4 y Tabla 4.5, respectivamente. En estas tablas, la fila *DT* muestra la desviación típica obtenida en cada método y la fila *Media* la precisión media obtenida. Las columnas *Etapa1*, *Etapa 2*, etc. indican la etapa, es decir, la cantidad de información de la que se ha partido para hacer la aproximación. Observando las tablas, puede verse como la precisión mayor se obtiene cuando el experto da toda la información, pero también puede verse como evoluciona cada modelo con respecto a la cantidad de información. Como muestran los datos, poca información, en ocasiones, no es suficiente, sin embargo ,mucha, puede dar lugar a que los resultados no sean adecuados.

La Tabla 4.2, muestra los resultados obtenidos con la interpolación mediante funciones splines. Se puede observar como con la base BUPA Liver disorder, el mejor resultado se obtiene en la *Etapa 2*, esto es con el 30 % de la información, éste resultado empeora un poco cuando tenemos el 50 % de la información. Esto era esperado puesto que la interpolación no es un método estable para grandes cantidades de información. También puede verse como en los experimentos con la base German credit ocurre exactamente igual y cuando crece demasiado la cantidad de información, empeoran los resultados, al igual que con Glass idetification, pero no con Wine recognition. Este último no es un resultado determinante, puesto que como ya se comentó en el Capítulo 3, este problema no encaja totalmente en la clase *Problemas Con Riesgo* y es justo dentro de esta clase donde nuestro modelo obtiene los mejores resultados.

La Tabla 4.3, muestra los resultados obtenidos usando aproximación por mínimos cuadrados. A rasgos generales puede decirse que son buenas aproximaciones. Pero, sin embargo, con este método puede verse como al aumentar la cantidad de información aumenta también la precisión. Esto es debido a que la aproximación es un método más estable, puesto que es más flexible que la interpolación con respecto a sus restricciones, ya que no está obligado a cumplirlas con exactitud. En la Tabla 4.4 pueden verse los resultados obtenidos con Wang y Mendel, método con el que menor precisión se ha obtenido.

Finalmente en la Tabla 4.5, aparecen reflejados los resultados que se obtienen cuando no se dispone de ninguna información de partida. Los resultados aproximan bastante bien a los del experto, esto es debido a que en la base de casos solo se guardan casos con solución correcta, y es esta información la que se utiliza para aproximar los valores de la función IRL. Estos resultados ayudan a confirmar la propiedad de que este concepto permite aprendizaje.

4.4.3. Análisis de los resultados experimentales

Para finalizar y en base al trabajo realizado se expondrán las razones que servirán como criterios de selección para decidir que método resulta adecuado en función de la situación. Para ello se sigue la misma división que se hizo al comienzo de este capítulo:

Tabla 4.2: Resultados obtenidos en cada etapa bajo la *Hipótesis 1*

		Etapa1	**Etapa2**	**Etapa3**	**Etapa4**
BUPA Liver	Media	*0.5684*	*0.6331*	*0.6149*	*0.7073*
	DT	0.0176	0.0254	0.0211	0.0093
German credit	Media	*0.7078*	*0.7200*	*0.7172*	*0.7626*
	DT	0.0018	0.0033	0.0028	0.0100
Wine Recognition	Media	*0.8974*	*0.8637*	*0.8812*	*0.9500*
	DT	0.0131	0.195	0.0148	0.0097
Glass Identification	Media	*0.9094*	*0.9151*	*0.9003*	*0.9661*
	DT	0.0042	0.0046	0.0075	0.0088

Tabla 4.3: Resultados obtenidos en cada etapa bajo la *Hipótesis 2*

		Etapa1	**Etapa2**	**Etapa3**	**Etapa4**
BUPA Liver	Media	*0.5843*	*0.6116*	*0.6413*	*0.7073*
	DT	0.0101	0.0093	0.0118	0.0093
German credit	Media	*0.7203*	*0.7180*	*0.7184*	*0.7626*
	DT	0.0024	0.0015	0.0021	0.0100
Wine Recognition	Media	*0.9306*	*0.9324*	*0.0.9424*	*0.9500*
	DT	0.0070	0.0046	0.0046	0.0097
Glass Identification	Media	*0.9118*	*0.9508*	*0.9580*	*0.9661*
	DT	0.0024	0.0021	0.0035	0.0088

1. Cuando un experto está siempre a nuestra disposición.
2. Cuando el experto tan solo da información parcial del problema.
3. Cuando no se disponemos de ninguna información, es decir, ha sido imposible disponer de un experto.

La primera situación ya quedó completamente resuelta en el capítulo anterior. Para la situación 3, en la que no se dispone de información alguna, el método más adecuado es el riesgo probabilístico ya que ha sido diseñado especialmente para resolver esta situación. Tras los experimentos se ha visto que es capaz de obtener un buen comportamiento, aproximando al mismo nivel de otros métodos. No Obstante, no llega a superar a los demás cuando estos están bajo las condiciones óptimas. Los

Tabla 4.4: Resultados obtenidos en cada etapa bajo la *Hipótesis 3*

		Etapa1	**Etapa2**	**Etapa3**	**Etapa4**
BUPA Liver	Media	*0.5649*	*0.5958*	*0.5940*	*0.7073*
	DT	0.0056	0.0040	0.0032	0.0093
German credit	Media	*0.7080*	*0.7090*	*0.7010*	*0.7626*
	DT	0.0172	0.0170	0.0130	0.0100
Wine Recognition	Media	*0.8775*	*0.8900*	*0.8650*	*0.9500*
	DT	0.0818	0.0805	0.0813	0.0097
Glass Identification	Media	*0.8190*	*0.8494*	*0.7714*	*0.9661*
	DT	0.0030	0.0023	0.0010	0.0088

Tabla 4.5: Resultados obtenidos en cada etapa bajo la *Hipótesis 4*

		–	**Etapa4**
BUPA Liver	Media	*0.6293*	*0.7073*
	DT	0.1125	0.0093
German credit	Media	*0.7290*	*0.7626*
	DT	0.0186	0.0100
Wine Recognition	Media	*0.9187*	*0.9500*
	DT	0.0562	0.0097
Glass Identification	Media	*0.9380*	*0.9661*
	DT	0.0565	0.0088

resultados muestran que la información almacenada en la base de casos, aunque es limitada para poder llegar a competir con el conocimiento de un experto, sí ofrece una valiosa información debido principalmente a que toda ella proviene de casos resueltos con éxito.

Queda por analizar la situación 2, tal vez sea la más complicada, ya que decidir qué hacer cuando solo se tiene parte de la información no depende de nosotros, sino como se ha visto, será la situación concreta la que marque el método más adecuado. Si se dispone de muy poca información numérica sería adecuado utilizar cualquier modelo. Sin embargo, en esta situación e independientemente de la procedencia de la información (y de cualquier criterio que aconseje el experto) los modelos más adecuados serían Lagrange y Hermite. Para pocos puntos estos aproximan mejor incluso

que las funciones spline. Por otro lado, se ha visto como los modelos de aproximación dan mejores resultados cuanta más información de partida se les proporciona, luego no resultaría adecuado utilizarlos en esta situación. Tampoco sería adecuado utilizar Wang y Mendel, ya que carece de sentido con poca información construir una base de reglas, puesto que ésta apenas cubrirá parte del conocimiento necesario para obtener una aproximación razonable a la realidad.

Si la cantidad de información aumenta, los primeros modelos a descartar son Lagrange y Hermite, sin embargo podrían emplearse todos los demás. El más indicado será interpolación mediante funciones splines cuando la información sea datos muestrales (que procedan de experimentos o muestras reales) y/o el experto informe de que esa información ha de cumplirse con exactitud. En este caso los modelos de aproximación y el modelado difuso de los datos son opciones posibles, pero no las más adecuadas, puesto que simplemente aproximarán la información en una situación que requiere exactitud. En esta misma situación, si la cantidad de información crece demasiado, será más conveniente elegir entre aproximación y Wang y Mendel ya que la pérdida de estabilidad de la interpolación hace perder en la precisión final de la aproximación.

Cuanta más información se tenga serán los métodos de aproximación y el de modelado difuso de datos los más adecuados, con independencia de información extra acerca del rigor con que debe cumplirse. Los experimentos muestran como, cuanta más información de partida usa el método de aproximación, mayor precisión obtiene. Esto se debe a que la aproximación tiene restricciones más débiles, puesto que no está obligada a cumplir con exactitud cada punto. En esta situación, aplicar Wang y Mendel también resulta adecuado y aunque los resultados muestran que su comportamiento es poco predecible, esto es debido a su alta flexiblidad. Llegados a este punto, esta flexibilidad puede suponer tanto una ventaja como una desventaja. Por un lado tener parámetros libres como el número de etiquetas difusas o el tipo de función de pertenencia pueden ayudar a obtener muy buenos resultados, pero obligan a contar con la ayuda del experto o hacer pruebas de ajuste.

Finalmente, puede decirse que si la información de la que se parte tiene el suficiente peso (según el experto) como para estar obligados a cumplirla con exactitud, se aconseja que si la cantidad de información es muy grande se utilice interpolación mediante funciones splines, o si la información fuese muy poca (menos de 6 puntos por atributo) se aconseja Lagrange como mejor aproximación. Si se dispone de una

cantidad razonable de información se aconseja aproximación o Wang y Mendel. Entre ellos se debe elegir en función de la ayuda del experto. Si este ayuda a fijar los parámetros, Wang y Mendel sería un método muy adecuado, si no es el caso, mejor utilizar aproximación, método al que avalan los buenos resultados obtenidos.

4.5. Resumen y conclusiones

La Información de Riesgo, como se vio en el capítulo anterior, nos ayuda a recuperar no sólo el caso más similar, sino también el caso más conveniente para resolver el problema. Pese a funcionar bien y proporcionar buenos resultados presenta un inconveniente: su dependencia constante del experto. Para resolver este problema a lo largo de este capítulo se han propuesto y analizado diferentes métodos para predecir el valor de la función IR.

El estudio de los métodos quedó marcado por dos situaciones principales: cuando se contaba parcialmente con ayuda del experto, (se tenía información de partida) y cuando no se disponía de información alguna. Para resolver la primera situación, se tomó de los métodos existentes en la literatura algunos de los más comúnmente utilizados como son los modelos de interpolación, aproximación y modelado difuso de datos. Para dar solución a la otra situación se desarrolló un método propio, que aproxima el riesgo tomando la información que proporcionan los casos almacenados en memoria.

Para obtener conclusiones sobre el comportamiento de estos método y saber en función de la situación qué método resultaba más adecuado, se experimentó cada método bajo cuatro hipótesis distintas marcadas por etapas. Las etapas indicaban la cantidad de información de partida. Tras obtener los resultados se pudo comprobar, como la interpolación es la mejor aproximación cuando se dispone de poca información, mientras que la aproximación resulta muy apropiada si la cantidad de ejemplos de partida es muy amplia. Wang y Mendel proporcionan buenas aproximaciones, pero con la desventaja de que para conseguirlas hay que ajustar cada uno de sus parámetros. También el riesgo probabilístico da una buena aproximación si nos encontramos sin ninguna información de partida.

Capítulo 5

Funciones de Ganancia y Pérdida para Recuperación en RBC

En este capítulo se profundiza en el estudio de las consecuencias que conlleva elegir una u otra solución a la hora de recuperar un caso. Ya que el tomar esa decisión existen factores cruciales que deberían ser considerados, como son la ganancia y pérdida potencial asociadas a la solución escogida. Por tanto, a lo largo de capítulo se introducen los conceptos de Ganancia y Pérdida de cada solución [27]. Estos conceptos, que serán definidos como funciones, miden la ganancia y la pérdida de aplicar una solución con éxito. Además, ayudarán a elegir la solución que reporte el mayor beneficio de entre todas las posibles soluciones; por tanto, en caso de cometer un error se sufrirá la mínima pérdida, por lo que se conseguirá siempre obtener el mayor beneficio posible.

5.1. Introducción

Como ya se ha comentado a lo largo de esta memoria, la mayoría de técnicas usadas en recuperación, solo tienen en cuenta a la hora de decidir qué caso recuperar, la similitud entre atributos y la importancia o peso que cada atributo tiene en el problema. En el capítulo 3, se estudió como el éxito del caso recuperado está influenciado por lo adecuada o no que es la solución asociada a dicho caso para resolver el problema. Esta *Adecuación* de la solución se medía gracias a la aplicación *Información de Riesgo Local* y se hacía atributo por atributo. Es más, introduciendo este conjunto de aplicaciones como una nueva información del problema, se obtuvieron

buenos resultados experimentales. Ahora, la intención es dar otro paso y analizar con profundidad la idea de que el éxito del caso recuperado queda influenciado por lo adecuada o no que sea la solución asociada. Para ello se introducen dos nuevos conceptos que son la *Ganancia* y la *Pérdida* de cada solución. Considerando estos conceptos antes de recuperar un caso, se tendrá en cuenta cómo de adecuada es la solución asociada, pero ahora en función de la ganancia y la pérdida que ésta produce. Esto resultará muy útil puesto que en problemas reales la ganancia y la pérdida no son complementarias ($Ganancia \neq -\ Pérdida$).

A continuación se ilustran estos conceptos a través de un ejemplo. El contexto del ejemplo será la medicina. El Razonamiento Basado en Casos es considerado una técnica apropiada para soporte al diagnóstico en medicina [18, 55, 72, 197, 5]. Esto se debe principalmente a que la medicina es uno de los campos donde es muy útil tener una memoria de casos reales y RBC proporciona precisamente ese conocimiento. La situación es la siguiente, hay cuatro pacientes almacenados en memoria de los que se conocen los síntomas que presentaban en el momento en que llegaron a urgencias, estos síntomas son: localización del dolor, otros síntomas como náuseas, vómitos, etc., fiebre, si presenta o no pérdida de apetito y su edad. Las posibles enfermedades relacionadas con esos síntomas son: Apendicitis, Gastritis y Flato. Llega un paciente a urgencias, al que se llamará (*Paciente Nuevo*) del que se conoce su enfermedad que es Apendicitis, pero se quiere comprobar qué diagnosticaría un sistema RBC. La Tabla 5.1, muestra los pacientes o casos, los atributos o síntomas, los pesos o importancia de cada síntoma y las soluciones o enfermedades asociadas.

Tabla 5.1: Base de casos y pesos

Atributos	**Pesos(ω_i)**	**Pacien. 1**	**Pacien. 2**	**Pacien. 3**	**Pacien. 4**	**Pacien. Nue.**
Localiza dolor	0.91	Flanco derecho	Flanco derecho	Epigastrio	Epigastrio	Flanco derecho
Otros síntomas	0.78	Vómitos	Mareos	Náuseas	Ninguno	Náuseas
Fiebre	0.60	38.7	37.5	36.8	38.2	37.8
Pérdida apetito	0.40	Si	Si	No	Si	Si
Edad	0.20	11	35	20	25	14
DIAGNÓSTICO		APENDICITIS	GASTRITIS	FLATO	GASTRITIS	¿?

Para tener un primer diagnóstico, se calcula la similitud de los casos en memoria con respecto al caso actual, *Paciente Nuevo*, Este proceso se hará en dos pasos, primero se calcula la similitud local entre atributos con las medidas dadas en

Tabla 5.2: Similitud local y global de los casos en relación al caso Paciente Nuevo

	Loc. dolor	**Otros sínt.**	**Fiebre**	**Pérd. apetito**	**Edad**	**Sim. Glob.**
Paciente 1	1	0	0.5263	1	0.8750	0.6231
Paciente 2	1	0	0.8421	1	0.1250	**0.6367**
Paciente 3	0	1	0.4736	0	0.7500	0.4201
Paciente 4	0	0	0.7894	1	0.5417	0.3397

Ecuación 3.1 y Ecuación 3.2 dependiendo si el atributo es discreto o continuo y luego se calcula la similitud global del caso como la media ponderada entre la similitud local de cada atributo y su peso correspondiente, según Ecuación 3.4.

En la Tabla 5.2 pueden verse los resultados obtenidos. El caso más similar a *Paciente Nuevo* es *Paciente 2* y por lo tanto, la solución que se debería aplicar, o lo que es lo mismo, la enfermedad que se le diagnostica es *Gastritis*. Sin embargo, esta solución no es apropiada puesto que su verdadera enfermedad es *Apendicitis*. Además, en este caso, diagnosticar *Gastritis* pone la vida del paciente en peligro. Esto es debido a que el sistema solo ha tenido en cuenta para recuperar el caso la similitud entre atributos y el peso de cada uno de ellos, perdiendo así, información importante del problema, como son las consecuencias que conlleva asociadas el aplicar una u otra solución.

Veamos, en detalle, qué tipo de información se pierde. La verdadera enfermedad del paciente es *Apendicitis*. Por tanto, si se decide diagnosticar gastritis o flato, la pérdida potencial es altísima ya que se pone en peligro la vida del paciente al no proporcionarle el tratamiento adecuado. La falta de dicho tratamiento puede producir una inflamación del perineo, la cual, incluso, puede llegar a provocarle la muerte. Por otro lado, si se diagnostica apendicitis, hay una ganancia altísima, puesto que éste recibirá el tratamiento correcto y su vida no correrá ningún peligro. Suponiendo ahora que la verdadera enfermedad del paciente es Gastritis, si se le diagnostica flato la perdida potencial es media. Aunque no se trate correctamente su enfermedad, y pese a que pudiera empeorar, ésta no conllevaría ninguna consecuencia peligrosa para el paciente. Si se le diagnostica apendicitis, sin tenerla, la perdida es alta, ya que aplicar el tratamiento asociado a esta enfermedad implicaría someter al paciente a una intervención quirúrgica sin necesidad. Por otro lado existirá ganancia alta, (aunque no altísima como en el caso anterior en el que el paciente tenía apen-

dicitis y se le diagnosticaba apendicitis) si se le diagnostica gastritis puesto que el tratamiento correcto resolvería el problema y en ningún momento correría peligro la vida del paciente. De forma análoga, se actuaría en el caso en que la verdadera enfermedad del paciente fuese flato.

Por lo tanto, parece interesante conseguir que el sistema considere toda esa información antes de recuperar un caso. Ya que equivocarse en los casos cuyo coste sea menor es mejor que equivocarse en los casos donde el coste sea altísimo. Por esto, uno de los objetivos de esta memoria es tratar de minimizar el coste de cada decisión, en cada recuperación que realice el sistema. Donde el coste, al que más adelante se llamará *Beneficio* de la solución, será definido como la diferencia media de la ganancia menos la pérdida. Tomando en cuenta estas razones, es clara la necesidad de introducir los conceptos de *Ganancia* y *Pérdida* de una solución en la recuperación del caso. Por tanto, a lo largo de este capítulo, se propone un método de recuperación que haga uso de la información proporcionada por dichos conceptos. Ambos serán definidos como funciones, cuyo objetivo será medir la pérdida y la ganancia que existe en cada decisión tomada. Concretamente, el objetivo de este capítulo es minimizar el coste de cada decisión. Para conseguir esto se introduce la función *Beneficio de que la función sea usada con éxito* esta función será definida a través de las funciones *Pérdida* y *Ganancia*.

El resto de este capítulo se organiza en 4 secciones. En la Sección 2 se introducen los conceptos de Ganancia, Pérdida y Beneficio de la solución considerada. En la Sección 3, se presenta el sistema de inferencia difuso compuesto por 19 reglas, que son las encargadas de asignar la similitud global basándose para ello en el beneficio potencial de la solución. En la Sección 4 se evalúa el modelo utilizando para ello bases de datos públicas procedentes del UCI-*machine learning repository*, los resultados obtenidos será comparados con los de conocidos modelos. Finalmente, las conclusiones del capítulo serán expuestas en la Sección 5.

5.2. Pérdida, Ganancia y Beneficio de la solución

Como se ha visto en el ejemplo de la Sección 1, una vez calculada la similitud entre los casos usando únicamente la información proveniente de la base de casos, se pierde una parte muy importante de la información que proporciona el problema. Luego, surge la necesidad de combinar la información que aportan los casos almacenados en memoria con otros aspectos relevantes del problema para así poder tomar

la mejor decisión. Por tanto, a partir de ahora, cuando se recupere un caso se tendrán en cuenta dos nuevos factores: la probabilidad de que una solución sea usada correctamente y las consecuencias asociadas a esa decisión, esta última información será dada por un experto. En esta sección, se definen los conceptos necesarios para conseguir este objetivo. Estos conceptos ayudarán, tanto a tomar la mayor cantidad de información posible de la base de casos, como a introducir en el problema todo tipo de información relevante que ayude a encontrar la solución más adecuada. A continuación se analizará cada uno en profundidad.

5.2.1. Probabilidad de que una solución sea usada con éxito

Para calcular la probabilidad de que una solución, S, sea usada con éxito, se pueden plantear las siguientes opciones:

1. Que el experto dé esa información (seria la situación ideal).
2. Conocer la probabilidad inicial de ocurrencia de cada solución, y luego, mediante reiteradas aplicaciones del Teorema de Bayes [14], calcular la probabilidad de ocurrencia de cada solución condicionada al nuevo caso.

En esta memoria se utilizará la segunda opción para estimar la probabilidad. Para calcularla haremos uso de la información almacenada en la base de casos.

5.2.1.1. Probabilidad Inicial o Probabilidad a Priori

Cada vez que llegue un nuevo caso se le asignará una probabilidad inicial por cada posible solución S. Esta probabilidad inicial se calculará usando la información de la base de casos y la Definición 5.1.

Definition 5.1. Sea S una de las soluciones almacenadas en memoria, la *Probabilidad Inicial* de la solución S, $P_0(S)$, se define como:

$$P_0(S) = \frac{\textit{Número de casos en memoria cuya solución es } S}{\textit{Número total de casos en memoria}}$$

Aspectos importantes a tener en cuenta sobre esta asignación de *Probabilidad Inicial*:

- La probabilidad inicial, $P_0(S)$, es la probabilidad asignada a un caso del que no se tiene ninguna información.

- Si el caso es finalmente almacenado en la base de casos, la probabilidad inicial $P_0(S)$, aumentará, puesto que habrá un caso más cuya solución correspondiente sea la solución S. En este sentido se puede hablar de aprendizaje.

- En contra, esta definición de probabilidad, en total ausencia de un experto, presenta el problema de ser demasiado dependiente de la base. Lo que no impide que a medida que el sistema vaya aumentando el número de casos, esta probabilidad converja a la verdadera probabilidad.

5.2.1.2. Probabilidad Condicionada o a Posteriori

Tan pronto como se conocen los atributos del caso, se produce una modificación en las probabilidades de las soluciones asociadas. Por ello, resulta necesario actualizar las probabilidades iniciales anteriores con la información que cada atributo aporta sobre el caso. Es aquí donde la definición de probabilidad condicionada y el Teorema de Bayes [14] juegan un papel importante. Ahora, lo que interesa es conocer, la probabilidad de la solución que se está considerando condicionada a los valores concretos de los atributos del caso. Al utilizar el Teorema de Bayes para actualizar las probabilidades iniciales, es necesario tener en cuenta que:

- Cada atributo puede aumentar o disminuir la probabilidad de que esa solución sea usada con éxito.

- El conocimiento *a priori* se combina con el conocimiento actual.

- La información obtenida tiene en cuenta todas las posibles combinaciones.

Veamos ahora cómo utilizando estas dos herramientas se calcula la probabilidad de una solución conocido el caso nuevo, es decir, la probabilidad de la solución S condicionada a los valores de los atributos del nuevo caso, $P(S/Caso\ Nuevo)$. Sean $S_1, \ldots, S_m$ las m soluciones distintas almacenadas en la base de casos, dónde cada caso tiene n atributos $A_1, \ldots, A_n$. Entra un nuevo caso, $Caso\ Nuevo = (A_1 = a_1, \ldots, A_n = a_n)$ donde $a_1, \ldots, a_n$ son los valores concretos que para *Caso Nuevo* toman cada uno de sus atributos. Ahora, se actualiza la probabilidad inicial de ocurrencia de cada solución, con la información que cada valor concreto del atributo de *Caso Nuevo* aporta al problema.

El procedimiento de forma detallada es el siguiente: se fija la solución, S_i y se calcula su probabilidad inicial asociada, $P_0(S_i)$, utilizando para ello la Definición 5.1.

Luego, esta probabilidad será actualizada en función de la información que aportan los valores de los atributos del caso nuevo mediante reiteradas aplicaciones de la definición de probabilidad condicionada y del Teorema de Bayes, hasta obtener $P(S_i/Caso\ Nuevo)$. Para ello, se introducen los atributos uno a uno y se calculan sus respectivas probabilidades. Para el atributo 1 se calcula $P_1(S_i) = P(S_i/A_1 = a_1)$, para el atributo 2, $P_2(S_i) = P(S_i/A_1 = a_1 \cap A_2 = a_2)$, y así hasta el n-*ésimo*, $P_n(S_i) = P(S_i/A_1 = a_1 \cap, \ldots, \cap A_n = a_n) = P_n(S_i/Caso\ Nuevo)$. Cada una de estas probabilidades se obtiene aplicando la definición de probabilidad condicionada y el teorema de Bayes. Las frecuencias de ocurrencia de cada valor de cada atributo serán tomadas de la base de casos. Por tanto, P_1, quedará

$$P_1(S_i) = P(S_i/A_1 = a_1) = \frac{P(A_1 = a_1/S_i) \cdot P_0(S_i)}{P(A_1 = a_1/S_1) \cdot P_0(S_1) + \ldots + P(A_1 = a_1/S_m) \cdot P_0(S_m)}$$

De forma análoga se calcula el resto de las probabilidades

$$P_j(S_i) = P(S_i/A_j = a_j \cap \ldots \cap A_1 = a_1) =$$

$$= \frac{P(A_j = a_j/S_i) \cdot P_{j-1}(S_i)}{P(A_j = a_j/S_1) \cdot P_{j-1}(S_1) + \ldots + P(A_j = a_j/S_m) \cdot P_{j-1}(S_m)} =$$

$$= \frac{P(A_j = a_j/S_i) \cdot P_0(S_i/A_{j-1} = a_{j-1} \cap \ldots \cap A_1 = a_1)}{P(A_j = a_j/S_1) \cdot P(S_1/X) + \ldots + P(A_j = a_j/S_m) \cdot P(S_m/X)}$$

donde $X = (A_{j-1} = a_{j-1} \cap \ldots \cap A_1 = a_1)$ y cada $P(A_j = a_j/S_i)$ se calcula como sigue:

$$P(A_j = a_j/S_i) = \frac{\text{Número de casos en la base de casos cuyo atributo } A_j \text{ verifica D}}{\text{Número de casos en la base de casos cuya solución asociada es } S_j}$$

donde D=$\{A_j = \nu \ / \ | \ \nu - a_j| \leq \alpha,\ \alpha \in \mathbb{R}\ \}$. Este es el conjunto de atributos que verifica $| \ \nu - a_j| \leq \alpha$, es decir, los atributos que distan de a_j menos que α y

donde ν en cada caso es el valor del atributo del caso con que comparamos y a_j el verdadero valor del atributo, es decir, el valor del atributo del caso nuevo.

Y así sucesivamente hasta calcular $P_n(S_i/\ Caso\ Nuevo)$. Hay que notar que tanto la definición de probabilidad inicial como la de probabilidad condicionada pueden usarse para atributos categóricos o atributos continuos, ya que estas definiciones cuentan el número de atributos que cumplen unas ciertas restricciones y por tanto no importa el tipo de atributo.

5.2.2. Beneficio, Pérdida y Ganancia de que una solución sea usada con éxito

Ya se ha comentado que uno de los objetivos planteados en esta memoria es recuperar no sólo el caso más similar, sino el caso que además proporcione mayor beneficio. Para determinar dicho beneficio se medirá la ganancia y la pérdida asociada a cada solución S, almacenada en el sistema. Para ello, a lo largo de esta sección se presentan dos nuevas funciones llamadas *Pérdida* y *Ganancia* asociada a cada solución.

5.2.2.1. Funciones Pérdida y Ganancia

Sea D el conjunto de todas las posibles soluciones del problema, d una solución particular y Θ el espacio de incertidumbre asociado. Se sabe que en cada decisión tanto la ganancia como la pérdida asociadas son elementos claves, por lo que poder medirlas facilitaría mucho el trabajo. Si se toma una solución particular d_i del problema, del que además se sabe que s_j es la verdadera solución, esto generará una pérdida a la que se notará $L(d_i, s_j)$. Por tanto, se define la *Función pérdida* $L(d, s)$, para cada $(d, s) \in D \times \Theta$ como la pérdida en que incurre el problema al utilizar la solución d para resolverlo, cuando la verdadera solución es s. Por simplicidad, solo se considerarán las funciones pérdida que cumplan $L(d, s) \geq -K > -\infty$. Esta condición deberá ser verificada por cualquier función pérdida de interés, ya que no tiene sentido realizar cálculos en un problema con pérdidas infinitas. La incorporación de la función pérdida en el análisis estadístico fue propuesta por Abraham Wald [169]. De forma similar, si una solución particular, d_i, es escogida y s_j es la verdadera solución, entonces el problema conlleva una ganancia a la que notaremos $G(d_i, s_j)$. Por tanto, se define la *Función ganancia* $G(d, s)$ para todo $(d, s) \in D \times \Theta$, como la ganancia en que incurre el problema al tomar la solución d cuando la verdadera

Tabla 5.3: Matriz de pérdidas

$L(d_i, s_j)$	**Apendicitis**	**Gastritis**	**Flato**
Apendicitis	*0*	*Alta*	*Alta*
Gastritis	*Muy Alto*	*0*	*Media*
Flato	*Muy Alto*	*Media*	*0*

solución es s. Esta función verifica $0 \leq G(d, s) < \infty$.

A modo ilustrativo, véase como se construye la función pérdida para el ejemplo presentado en la Introducción, dónde $\Theta = \{s_1 = Apendicitis, s_2 = Gastritis, s_3 = Flato\}$ y $D = \{d_1 = Apendicitis, d_2 = Gastritis, d_3 = Flato\}$. Para asignar la pérdida, se han usado etiquetas lingüísticas en lugar de valores numéricos para que su compresión sea sencilla. El tipo de valores que se asigne a las funciones ganancia y pérdida, no implica ninguna restricción, puesto que conocidas las funciones de pertenencia, se puede hacer cualquier tipo de cálculo sin problema. La Tabla 5.3 muestra todos los valores, véase a continuación como se han asignado algunos de ellos detalladamente. *L(Gastritis, Apendicitis)= Muy Alto* esto representa la pérdida en que incurre el problema cuando a un paciente se le diagnostica gastritis, siendo su verdadera enfermedad apendicitis, por la misma razón expuesta en la Sección 1 de este capítulo el valor es *muy alto*. En el caso de la función ganancia el procedimiento es el mismo.

Cuando tanto Θ como D son espacios finitos la función pérdida suele ser representada como una matriz llamada *Matriz de Pérdidas*. Normalmente los valores del espacio D se colocan en la primera fila de la matriz y los del espacio Θ en la primera columna. La función pérdida no debería ser construida de esta forma, ya que cada problema debería tener su propia función pérdida bien definida. El concepto de pérdida es difícil de modelar y por esta razón, en la literatura se pueden encontrar muchas investigaciones que discuten la construcción y existencia de este tipo de funciones véase [48, 58, 140, 7]. Otros métodos de construcción más sofisticados son el conocido *backwards* [15] o [84]. En teoría de la decisión cuando en un problema no se dispone de información muestral asociada al experimento, se le llama "problema sin datos", para este tipo de problemas la representación de la función pérdida como matriz es aceptada. Los problemas tratados en esta memoria son considerados como

"problemas sin datos", debido a que la mayor parte de la información extra usada para resolverlos es proporcionada por el experto y no se dispone de información muestral procedente de experimentos estadísticos. Luego se aceptarán las matrices de pérdidas y ganancias como representación de dichas funciones. Hay que resaltar que en cualquier otra clase de problemas en los que se tenga cualquier otro tipo de información, se podrá usar la función pérdida que mejor se ajuste al problema, como se ha visto en la bibliografía.

En Teoría de la Decisión y en Teoría de la Utilidad, la *Ganancia* es complementaria a la *Pérdida*, por tanto, conocida la ganancia, (llamada normalmente utilidad), fácilmente se obtiene la pérdida como $L(d,s) = -G(d,s)$. En problemas reales no siempre se pueden considerar estos dos conceptos complementarios. Por este motivo, en este trabajo, se han definido las funciones de ganancia y pérdida de forma independiente.

Como ya se ha mencionado anteriormente, cualquier toma de decisiones implica incertidumbre. Por tanto, la pérdida real, $L(d,s)$, nunca se conocerá con exactitud en el momento de tomar la decisión. Por esta razón, como método natural de proceder se ha considerado el cálculo de la *pérdida esperada*.

Definición 5.2. La esperanza matemática de una función $f(x)$, se define como

$$E[f(x)] = \sum_{i=1}^{\infty} f(x_i) \cdot p_i$$

dónde p_i es la probabilidad de ocurrencia de cada valor de la función $f(x_i)$.

Por lo tanto, la pérdida y la ganancia esperadas, serán definidas de la siguiente manera:

Definición 5.3. Sean $S_1, \ldots, S_m \in \Theta$, las posibles soluciones del problema. Se define la pérdida esperada, $E[L(S)]$, de una solución $d = S \in D$ como

$$E[L(S)] = E_{S_j \in \Theta}[L(d, s_j)] = L(S, S_1) \cdot P_n(S_1) + \ldots + L(S, S_m) \cdot P_n(S_m)$$

dónde $P_n(S_i) = P(S_i / Caso\ \ Nuevo)$ es la probabilidad condicionada (o probabilidad a posteriori) de la solución S_i.

De modo similar se define la ganancia esperada:

Definición 5.4. Sean $S_1, \ldots, S_m \in \Theta$, las posibles soluciones del problema. Se define la ganancia esperada, $E[G(S)]$, de una solución $d = S \in D$ como

$$E[G(S)] = E_{S_j \in \Theta}[G(d, s_j)] = G(S, S_1) \cdot P_n(S_1) + \ldots + G(S, S_m) \cdot P_n(S_n)$$

dónde $P_n(S_i) = P(S_i/Caso\ \ Nuevo)$ es la probabilidad condicionada (o probabilidad a posteriori) de la solución S_i.

5.2.2.2. Función Beneficio Objetivo de la Solución

En la literatura, se pueden encontrar distintos enfoques para toma de decisiones. Sin ir muy lejos, en estadística clásica existen varios procedimientos como: la máxima probabilidad [54], el estimador imparcial de riesgo, del que existen varias versiones como muestran los siguientes trabajos [130, 165, 41], la mínima varianza y el principio de mínimos cuadrados por nombrar algunos. También en teoría de la decisión hay principios que pueden ser usados con este objetivo. Los tres más importantes son: el principio de invarianza, el principio del riesgo de Bayes [14] y el principio minimax que presenta un enfoque diferente. La clave de este enfoque es que no conlleva comparaciones entre pérdidas, sino que selecciona una solución del espacio de soluciones utilizando la idea del equilibrio de Nash [118]. Ejemplos de aplicaciones de este enfoque pueden verse en los siguientes artículos [198, 177]. Sin embargo, en el modelo propuesto se utiliza el concepto de Beneficio para tomar decisiones. Este concepto se define como una función a partir de las funciones *L(d,s)* y *G(d,s)* a la que se llamará *Beneficio de que una solución sea usada con éxito.*

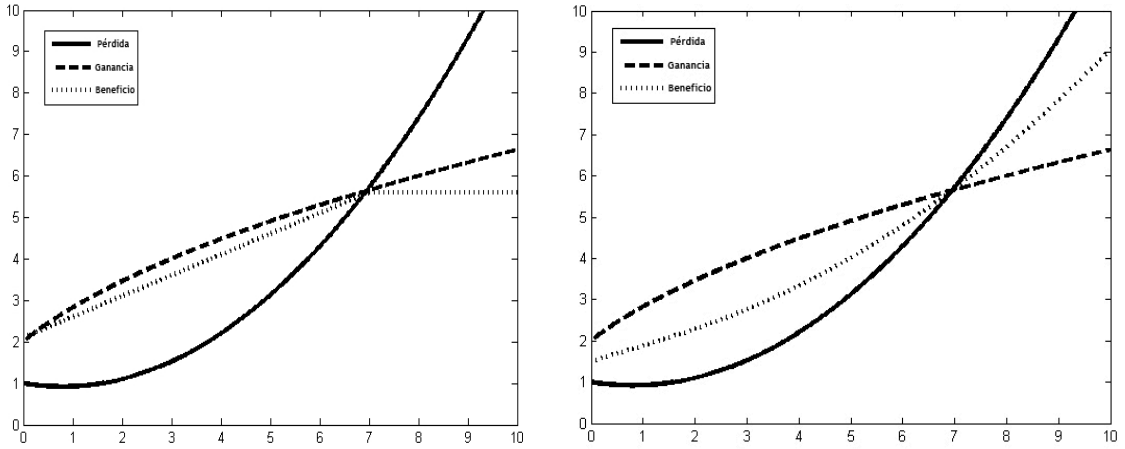
Definición 5.5. Se define la función Beneficio de que una solución sea usada con éxito como

$$B(d, s) = f(G(d, s), L(d, s))$$

y tal que,

I) sea creciente en la primera variable, es decir, creciente en la Ganancia

II) decreciente en la segunda variable, es decir, decreciente en la Pérdida.



(a) Modelo Conservador (b) Modelo utilizado en el modelo propuesto

Figura 5.1: Modelos de Beneficio

Como puede verse en la Definición 5.5, la elección de la función f ofrece una gran variedad de modelos alternativos de *Beneficio*. Desde modelos conservativos, dónde lo más importante es evitar pérdidas, hasta otros modelos más agresivos donde el único objetivo es obtener ganancias. La Figura 5.1 muestra ejemplos de estos modelos. En ella, pueden verse la función pérdida, la función ganancia y dos modelos diferentes de beneficio. La Figura 5.1(a) muestra un modelo conservador, ya que solo se está en situación de riesgo cuando las pérdidas son pequeñas y además el beneficio se estabiliza cuando las pérdidas crecen exponencialmente. La Figura 5.1(b) muestra el modelo de beneficio utilizado por el modelo propuesto, donde la función beneficio se define como: $B(d,s) = E_{S_j \in \Theta}[G(d,s_j) - L(d,s_j)]$, es decir, el beneficio es la media entre la ganancia y la pérdida. Se ha hecho esta elección de la función beneficio porque se tiene la intención de probar la nueva metodología desarrollada y no centrarse en resolver un problema particular. La función beneficio, que ha sido definida en Definición 5.5 es una definición subjetiva de beneficio, cuyo objetivo es considerar la participación del usuario. Esta función se estudiará en detalle en próximos trabajos.

5.3. Una nueva forma de recuperación usando el beneficio

Como se ha visto en las secciones previas de este capítulo, resulta interesante considerar el beneficio de cada solución antes de recuperar un caso. Por lo tanto, ahora no sólo la similitud y el peso de cada atributo son los que toman parte en el proceso de recuperación, sino también el beneficio que produce cada solución. Cuando se introdujo en la medida el riesgo local se eligió un sistema de inferencia difuso porque era útil y eficiente para tratar con información aproximada. Y también porque proporcionaba a la medida de similitud la información a veces lineal y a veces no lineal que el riesgo aportaba al problema. Por estas razones, para que el beneficio forme parte de la medida de similitud, también se ha elegido un sistema de inferencia difuso. Pese a esto, el beneficio presenta ciertas diferencias con respecto a la función información de riesgo:

El beneficio difiere del riesgo local en los siguientes aspectos:

1. El beneficio se asigna al caso entero y no atributo a atributo como se hacía con la información de riesgo.

2. El beneficio siempre puede ser asignado, incluso cuando no se tiene ayuda del experto.

3. Está incluido en la parte de adaptación del proceso, ya que los valores que toma el beneficio se actualizan con cada caso que entra en la base de casos.

4. Directamente afecta a la medida global y no afecta a la medida local.

5. Mientras el riesgo local es fijo, el beneficio se ajusta a sí mismo para minimizar el coste de cada decisión.

Para construir el sistema, se modifica la fórmula de la similitud global haciéndola depender del beneficio, $B(S)$, y de la similitud global estándar, $Sim(C^{Mem}, C^{Nue})$, como puede verse en la Ecuación 5.1:

$$Sim^{B}(C^{Mem}, C^{Nue}) = F(B(S), Sim(C^{Mem}, C^{Nue})) \tag{5.1}$$

La función $F(.)$ se obtiene implícitamente mediante un sistema de inferencia difuso y se calcula como la media ponderada de las salidas de todas las reglas. El

sistema de inferencia difuso usado en la Ecuación 5.1, al igual que el que se usó en la Ecuación 3.5, puede verse como una variación del modelo TSK [163]. Finalmente, el sistema de inferencia difuso contiene las siguientes 19 reglas:

Regla 1. Si B es *negativo-alto* y $Sim(C^{Mem}, C^{Nue})$ es *alto*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) - 0.1 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 2. Si B es *negativo-alto* y $Sim(C^{Mem}, C^{Nue})$ es *medio*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) - 0.2 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 3. Si B es *negativo-alto* y $Sim(C^{Mem}, C^{Nue})$ es *bajo*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) - 0.3 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 4. Si B es *negativo-medio* y $Sim(C^{Mem}, C^{Nue})$ es *alto*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) - 0.05 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 5. Si B es *negativo-medio* y $Sim(C^{Mem}, C^{Nue})$ es *medio*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) - 0.1 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 6. Si B es *negativo-medio* y $Sim(C^{Mem}, C^{Nue})$ es *bajo*, entonces V_i=$Sim(C^{Mem}, C^{New}) - 0.15 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 7. Si B es *negativo-bajo* y $Sim(C^{Mem}, C^{Nue})$ es *alto*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) - 0.025 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 8. Si B es *negativo-bajo* y $Sim(C^{Mem}, C^{Nue})$ es *medio*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) - 0.05 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 9. Si B es *negativo-bajo* y $Sim(C^{Mem}, C^{Nue})$ es *bajo*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) - 0.075 \cdot Sim(C^{Mem}, C^{Nue})$.

Regla 10. Si B es *cero*, entonces $V_i = Sim(C^{Mem}, C^{Nue})$.

Regla 11. Si B es *positivo-alto* y $Sim(C^{Mem}, C^{Nue})$ es *alto*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.3 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 12. Si B es *positivo-alto* y $Sim(C^{Mem}, C^{Nue})$ es *medio*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.2 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 13. Si B es *positivo-alto* y $Sim(C^{Mem}, C^{Nue})$ es *bajo*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.1 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 14. Si B es *positivo-medio* y $Sim(C^{Mem}, C^{Nue})$ es *alto*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.15 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 15. Si B es *positivo-medio* y $Sim(C^{Mem}, C^{Nue})$ es *medio*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.1 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 16. Si B es *positivo-medio* y $Sim(C^{Mem}, C^{Nue})$ es *bajo*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.05 \cdot Sim(C^{Mem}, C^{Nue})$.

Regla 17. Si B es *positivo-bajo* y $Sim(C^{Mem}, C^{Nue})$ es *alto*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.075 \cdot Sim(C^{Mem}, C^{Nue})$.

Regla 18. Si B es *positivo-bajo* y $Sim(C^{Mem}, C^{Nue})$ es *medio*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.05 \cdot Sim(C^{Mem}, C^{Nue})$

Regla 19. Si B es *positivo-bajo* y $Sim(C^{Mem}, C^{Nue})$ es *bajo*, entonces V_i=$Sim(C^{Mem}, C^{Nue}) + 0.025 \cdot Sim(C^{Mem}, C^{Nue})$

En las reglas, los términos lingüísticos (*negativo-alto, negativo-medio, negativo-bajo, cero, positivo-bajo, positivo-medio y positivo-alto*) se definen sobre $B(S)$, mientras que los términos (*bajo, medio, alto*) son definidos sobre la medida de similitud

$Sim(C^{Mem}, C^{Nue})$, V_i, $i = 1, \ldots, 19$ es la medida de similitud global ajustada según la regla í-*ésima* correspondiente.

El diseño de las reglas se hizo en dos pasos. Primero se tomaron los valores de la similitud global y del beneficio de cada solución, y según el valor concreto que tomasen se incrementaba o se bajaba la similitud global. Después se ajustaron los valores (0.1, 0.2, 0.3, ...) de los consecuentes de cada regla con la intención de obtener un buen resultado, para ello, primero se definieron como parámetros, y luego, utilizando la ayuda del experto y la experiencia del sistema, los valores con los que se obtuvieron mejores resultados fueron fijados. Por ejemplo, véase con detenimiento la *Regla 3*, esta regla se dispara cuando el beneficio es *negativo-alto* y la similitud toma el valor *bajo*. Cuando se dan estas dos condiciones se resta el 30 % sobre el resultado final de la similitud global para así ser capaces de evitar que el sistema recupere una solución con beneficio negativo, es decir, una solución que provoque pérdidas en el sistema. Condiciones similares son las que se dan en la *Regla 1*, pero en este caso al ser los casos muy similares, la cantidad que se resta es menor. La *Regla 10* se dispara cuando el beneficio es cero. En los casos de beneficio cero se deja al sistema que trabaje como siempre y no se modifica nada porque así se simplifica el trabajo y sólo se consideran los casos críticos, lo cual, es bastante congruente con la forma en que se está trabajando. Por otro lado, están las reglas en las que se aumenta la similitud global una cierta cantidad. Por ejemplo, la *Regla 12* se dispara cuando el beneficio es *positivo-alto* y la similitud toma el valor *medio*. Cuando se está bajo estas condiciones lo que se hace es subir la similitud global un 20 % para que el caso cuya solución tiene un beneficio positivo (lo que indica que apenas provocará pérdidas) tenga más posibilidades de ser recuperado.

Por último, la inferencia sobre el sistema, o equivalentemente el cálculo de $F(B(S), Sim(C^{Mem}, C^{Nue}))$ se hace en dos pasos. Primero se calcula la fuerza con que se dispara cada regla G_j, $j = 1, \ldots, k$, donde k es el número de reglas disparadas. La fuerza con que dispara cada regla se calcula mediante la agregación de los valores de verdad, es decir agregando los valores de la función de pertenencia de los términos lingüísticos utilizando para ello el operador "y" (en inglés *"and"*) al igual que hicimos en el capítulo 3 con la Información de Riesgo. Por ejemplo, la fuerza de disparo de la primera regla se obtiene según la Ecuación 5.2:

$$G_1 = \mu_{negativo-alto}(B(S)) \cdot \mu_{alto}(Sim(C^{Mem}, C^{Nue})) \tag{5.2}$$

Para todas las reglas con excepción de la *Regla 10*, la fuerza de disparo es calculada usando la Ecuación 5.2. Para la décima regla, que es una regla especial, la fuerza de disparo se calcula usando la Ecuación 5.3.

$$G_{10} = \mu_{cero}(B(S)) \tag{5.3}$$

Finalmente, se calcula la función $F(\cdot)$

$$Sim^B(C^{Mem}, C^{Nue}) = \frac{\sum_{j=1}^{k} G_j \cdot V_j}{\sum_{j=1}^{k} G_j} \tag{5.4}$$

dónde k es el número de reglas disparadas, V_j es la salida y G_j es la fuerza de disparo de la j-ésima regla, respectivamente.

5.4. Resultados experimentales

5.4.1. Descripción del problema: Casos de estudio

Para validar el sistema se aplica el método propuesto a cuatro bases de datos del UCI *machine-learning repository*. Las bases de datos han sido seleccionadas con el objetivo de validar la utilidad del beneficio en la recuperación de casos.

5.4.1.1. Los datos

Breast cancer data, este conjunto de datos proviene del Instituto de Oncología (Centro Médico Universitario) en Ljubljana, Yugoslavia. Fueron M. Zwitter y M. Soklic quienes dieron los datos para uso público. Originalmente contenía 286 casos. Cada caso tiene 9 atributos y representa información sobre tumores de mama en personas anónimas, los atributos son: la edad, si padece menopausia, premenopausia o nada, el tamaño del tumor, inv-nódulos, nódulo-cap, el estadío del tumor, si es la mama derecha o la izquierda, la zona de la mama en la que está y si ha sido o no irradiado. La base está dividida en dos clases: *Clase 1*, corresponde a los casos en que no se ha reproducido el tumor y *Clase 2* a los casos en que sí se ha reproducido. Solo se utilizaron 270 casos, para así construir subconjuntos de igual proporción y talla cara a la validación cruzada. Estos subconjuntos son seleccionados de forma aleatoria usando MATLAB 7.1. La Tabla 5.4 muestra los detalles de esta base de datos.

Tabla 5.4: Detalles de la base de datos usada en los experimentos (Breast cancer data)

Clase	Número de Casos	Porcentaje
Clase 1	190	70.37
Clase 2	80	29.63
Total	270	100.00

Tabla 5.5: Detalles de la base de datos usada en los experimentos (Pima Indians diabetes)

Clase	Número de Casos	Porcentaje
Clase 1	260	34.21
Clase 2	500	65.79
Total	760	100.00

Pima Indians Diabetes data set, este conjunto de datos pertenece al Instituto Nacional de Diabetes, Digestivo y Dolencias de Hígado y fue donado por Vincent Sigillito. Contiene 768 casos, con ocho atributos y dos clases. Cada caso representa un paciente, mujer, mayor de 21 años. Los atributos son: número de veces que ha estado embarazada, test oral de tolerancia a la glucosa, presión sanguínea, espesor de las capas de la piel del tríceps, cantidad de insulina, índice de masa corporal, edad y el valor de la función pedigrí de la diabetes. Solo se utilizaron 760 casos, para así construir subconjuntos de igual proporción y talla cara a la validación cruzada. Estos subconjuntos son seleccionados de forma aleatoria usando con MATLAB 7.1. La Tabla 5.5 muestra los detalles de esta base de datos.

BUPA liver disorder, para una explicación detallada véase Sección 3.4.1.1 del capítulo 3 y Tabla 3.5.

Heart diseases, donada por el Doctor R. Detrano del V.A. Medical Centrr. Fue construida originalmente en la Fundación Clinica Cleveland para el proyecto Stat-Log bajo el programa ESPIRIT del Comunidad Europea. Contenía 303 casos y 14 atributos contando la clase. Se eliminaron 7 casos para para así construir subconjun-

Tabla 5.6: Detalles de la base de datos usada en los experimentos (Heart diseases)

Clase	Número de Casos	Porcentaje
Clase 1	160	55.17
Clase 2	130	44.83
Total	290	100.00

tos de igual proporción y talla cara a la validación cruzada. Estos subconjuntos son seleccionados de forma aleatoria usando con MATLAB 7.1. La Tabla 5.6 muestra los detalles de esta base de datos.

5.4.1.2. Ejecución de los experimentos

Se realizará validación cruzada de 10 subconjuntos para cada base de datos. En cada validación se toma la base de datos entera y se divide en 10 subconjuntos excluyentes entre sí y con la misma distribución. Cada subconjunto se usa una vez como conjunto de prueba para ver los resultados que se obtienen al probarlo contra el conjunto que resulta de unir los otros nueve restantes subconjuntos. Se hicieron 10 validaciones completas, lo que supone un total de 100 repeticiones de cada experimento, para tener datos suficientes y poder así contrastar los resultados con t-test. Es el mismo procedimiento que se uso en el Capítulo 3 para probar la eficiencia de la Información de Riesgo.

Como objetivo se plantea evaluar los resultados obtenidos en los experimentos en términos de *precisión* y *beneficio*. La precisión mide la proporción de casos correctamente clasificados de entre todos los casos que han sido clasificados, es decir, la proporción de aciertos. Se ha elegido esta medida por ser una de las más usadas. Por otra lado, para medir el beneficio de una forma objetiva se tuvo en cuenta la opinión del experto. Éste informó de que la situación que mayor beneficio produce es cuando se clasifica como enfermo a un paciente que lo está, ya que esto permite aplicar el tratamiento correcto, directamente, sin pérdida de tiempo, evitando así que la vida del paciente corra peligro alguno. Por esta razón y para simplificar y hacer más objetiva la comparación con otros métodos, se considerará la proporción de casos de este tipo correctamente clasificados, y no el beneficio que el experto subjetivamente asigna a cada situación. Por tanto, se emplearán dos medidas más en el

estudio muy utilizadas en dominios médicos que son la *sensibilidad* y *especificidad* [182]. Sensibilidad mide la proporción de correctos positivos (CP) de entre todos los casos que representan a pacientes con la enfermedad, es decir, mide la proporción de casos que presentan la enfermedad y han sido correctamente clasificados. La especificidad mide la proporción de correctos negativos (CN) de entre todos los casos que representan a pacientes sanos, es decir, mide la proporción de casos que no presentan la enfermedad y han sido clasificados justo en esa clase.

En cada clasificación se consideraron los siguientes subconjuntos:

- CP, el paciente presenta la enfermedad y el diagnóstico es correcto.
- FP, el paciente presenta la enfermedad, pero el diagnóstico es incorrecto.
- CN, el paciente no sufre la enfermedad, y el diagnóstico es correcto.
- FN, el paciente no sufre la enfermedad y además el diagnóstico es incorrecto.

$$Precisión = \frac{CP + CN}{CP + FP + CN + FN}$$

$$Sensibilidad = \frac{CP}{CP + FP} \qquad Especificidad = \frac{CN}{CN + FN}$$

Sensibilidad es una medida muy usada y suele ser considerada como el porcentaje de detección de enfermedades, y está negativamente correlada con la especificidad. Por ejemplo, si el coste de clasificar mal un falso negativo es mayor que el de falso positivo, entonces la sensibilidad aumenta y la especificidad decrece. De acuerdo con el experto, clasificar mal un falso negativo es lo que conlleva las peores consecuencias cara al paciente, luego este estudio se centra en la precisión y la sensibilidad.

Tanto el Beneficio-RBC (BRBC) como el RBC estándar se implementaron en MATLAB 7.1. Para contrastar su eficiencia además, de con RBC estándar, se comparó con los siguientes modelos: J4.8, que es la implementación en WEKA del C4.5 (cuya versión comercial es el conocido C5.0) RBF Networks y k-NN con k=9. De todos estos algoritmos se usó la implementación que de ellos está disponible en WEKA [182]. Al igual que en el capítulo 3, para asegurar que exactamente los mismos subconjuntos que se generaron de forma aleatoria en MATLAB para la validación

cruzada eran usados en WEKA, se trabajó con esta herramienta en línea de comandos.

5.4.2. Resultados finales del estudio

Antes de comenzar los experimentos, se consultó con un experto para asignar los valores de las funciones ganancia y pérdida para cada base de datos. Después se asignaron los pesos, para ello se utilizaron los siguientes métodos de selección de pesos: Chi Square; Info Gain; Gain Ratio; y CfsSubsetEval, los cuales están disponibles en WEKA [182]. Tabla 5.7, Tabla 5.8, Tabla 5.9 y Tabla 5.11 muestran un resumen de los resultados obtenidos en precisión sensibilidad y especifidad para cada base de datos.

Tabla 5.7: Ranking de la precisión, sensibilidad y especificidad obtenida con Breast cancer

	1	**2**	**3**	**4**	**5**
Precisión	BCBR ($\frac{602}{810} = 0.7432$)	k-NN ($\frac{598}{810} = 0.7382$)	J4.8 ($\frac{593}{810} = 0.7320$)	RBF ($\frac{580}{810} = 0.7160$)	CBR ($\frac{485}{810} = 0.5987$)
Sensibilidad	BCBR ($\frac{127}{240} = 0.5292$)	CBR ($\frac{98}{240} = 0.4083$)	RBF ($\frac{86}{240} = 0.3583$)	k-NN ($\frac{85}{240} = 0.3541$)	J4.8 ($\frac{85}{240} = 0.3541$)
Especificidad	k-NN ($\frac{513}{570} = 0.9000$)	J4.8 ($\frac{508}{570} = 0.8912$)	RBF ($\frac{494}{570} = 0.8666$)	BCBR ($\frac{475}{570} = 0.8333$)	CBR ($\frac{387}{570} = 0.6789$)

Tabla 5.8: Ranking de la precisión, sensibilidad y especificidad obtenida con Pima Indian

	1	**2**	**3**	**4**	**5**
Precisión	J4.8 ($\frac{1734}{2280} = 0.7605$)	BCBR ($\frac{1717}{2280} = 0.7530$)	RBF ($\frac{1699}{2280} = 0.7451$)	k-NN ($\frac{1685}{2280} = 0.7390$)	CBR ($\frac{1571}{2280} = 0.6890$)
Sensibilidad	BCBR ($\frac{553}{780} = 0.7089$)	J4.8 ($\frac{485}{780} = 0.6217$)	RBF ($\frac{421}{780} = 0.5397$)	k-NN ($\frac{407}{780} = 0.5217$)	CBR ($\frac{400}{780} = 0.5128$)
Especificidad	RBF ($\frac{1278}{1500} = 0.8520$)	k-NN ($\frac{1278}{1500} = 0.8520$)	J4.8 ($\frac{1249}{1500} = 0.8326$)	CBR ($\frac{1171}{1500} = 0.7806$)	BCBR ($\frac{1164}{1500} = 0.7760$)

Tabla 5.9: Ranking de la precisión, sensibilidad y especificidad obtenida con BUPA Liver

	1	**2**	**3**	**4**	**5**
Precisión	J4.8 ($\frac{675}{1020} = 0.6617$)	RBF ($\frac{667}{1020} = 0.6539$)	BCBR ($\frac{650}{1020} = 0.6372$)	k-NN ($\frac{644}{1020} = 0.6313$)	CBR ($\frac{626}{1020} = 0.6137$)
Sensibilidad	BCBR ($\frac{519}{600} = 0.8650$)	J4.8 ($\frac{462}{600} = 0.7700$)	RBF ($\frac{462}{600} = 0.7700$)	k-NN ($\frac{450}{600} = 0.7500$)	CBR ($\frac{398}{600} = 0.6633$)
Especificidad	CBR ($\frac{228}{420} = 0.5428$)	J4.8 ($\frac{213}{420} = 0.5071$)	RBF ($\frac{205}{420} = 0.4880$)	k-NN ($\frac{194}{420} = 0.4619$)	BCBR ($\frac{131}{420} = 0.3119$)

Tabla 5.10: Ranking de la precisión, sensibilidad y especificidad obtenida con Heart disease

	1	**2**	**3**	**4**	**5**
Precisión	BCBR ($\frac{728}{870} = 0.8368$)	RBF ($\frac{713}{870} = 0.8195$)	k-NN ($\frac{706}{870} = 0.8115$)	CBR ($\frac{665}{870} = 0.7643$)	J4.8 ($\frac{657}{870} = 0.7551$)
Sensitivity	BCBR ($\frac{311}{390} = 0.7974$)	k-NN ($\frac{304}{390} = 0.7795$)	RBF ($\frac{302}{390} = 0.7744$)	CBR ($\frac{282}{390} = 0.7231$)	J4.8 ($\frac{265}{390} = 0.6795$)
Specificity	CBR ($\frac{417}{480} = 0.8688$)	RBF ($\frac{411}{480} = 0.8562$)	k-NN ($\frac{402}{480} = 0.8375$)	J4.8 ($\frac{392}{480} = 0.8166$)	BCBR ($\frac{383}{480} = 0.7979$)

Como puede verse en las tablas Beneficio-RBC es el que tiene la mayor media en precisión en los experimentos con las bases Heart disease y Breast cancer. Además, ocupa el segundo lugar en Pima Indian Diabetes con tan solo una pequeña diferencia con respecto al J4.8 y ocupa el tercer lugar en BUPA liver disorder. Sin embargo, en sensibilidad es el modelo, BRBC, quien consigue los mejores resultados en todos los experimentos. Como en este trabajo hemos supuesto que el coste de una mala clasificación cuando se trata de falsa ausencia (es decir, cuando clasificamos no enfermo a un paciente enfermo) es mayor que el opuesto (clasificar como enfermo a alguien que no lo está) podemos afirmar basándonos en estos resultados que se ha conseguido el objetivo propuesto de minimizar el coste de una mala clasificación. También en las tablas pueden verse los resultados obtenidos en especificidad. En esta medida BRBC ocupa el cuarto y el quinto puesto en los experimentos, pero con diferencias poco significativas. Se ha indicado la especificidad para mostrar como para conseguir el objetivo se han tenido que sacrificar algunos valores de especificidad.

Finalmente y para verificar que los datos son estadísticamente significativos, se contrastaron utilizando t-test con un nivel de confianza del 95 %. Tabla 5.11, Tabla 5.12, Tabla 5.13 y Tabla 5.14 muestran los p-valores obtenidos. Como muestran los resultados de los test existen diferencias significativas de nuestro método con respecto a todos los demás en sensibilidad, dando así más consistencia al objetivo planteado de minimizar el coste de una mala clasificación. También puede observarse que en precisión no existen diferencias significativas con respecto a ningún método.

Tabla 5.11: Resultados de los t-test al contrastar BRBC con cada modelo (Breast cancer)

	J4.8-BRBC	**RBF-BRBC**	***k*-NN-BRBC**	**CBR-BRBC**
Precisión	0.458	0.025	0.707	0.000
Sensibilidad	0.000	0.000	0.000	0.000
Especificidad	0.005	0.014	0.001	0.000

Tabla 5.12: Resultados de los t-test al contrastar BRBC con cada modelo (Pima Indian)

	J4.8-BRBC	**RBF-BRBC**	***k*-NN-BRBC**	**CBR-BRBC**
Precisión	0.425	0.289	0.097	0.000
Sensibilidad	0.000	0.000	0.000	0.000
Especificidad	0.000	0.000	0.000	0.722

Tabla 5.13: Resultados de los t-test al contrastar BRBC con cada modelo (BUPA liver)

	J4.8-BRBC	**RBF-BRBC**	***k*-NN-BRBC**	**CBR-BRBC**
Precisión	0.212	0.318	0.948	0.469
Sensibilidad	0.001	0.000	0.000	0.000
Especificidad	0.000	0.000	0.000	0.000

Tabla 5.14: Resultados de los t-test al contrastar BRBC con cada modelo (Heart disease)

	J4.8-BRBC	**RBF-BRBC**	***k*-NN-BRBC**	**CBR-BRBC**
Accuracy	0.000	0.030	0.062	0.000
Sensitivity	0.000	0.095	0.379	0.000
Specificity	0.006	0.281	0.045	0.001

5.5. Resumen y Conclusiones

A lo largo de este capítulo se ha profundizado en las consecuencias que conlleva asociada la toma de una decisión, en la recuperación de un caso, analizándolas en términos de Ganancia y Pérdida de la solución. Teniendo estas consecuencias en cuenta, se ha desarrollado un sistema de RBC que ofrece la posibilidad de considerar el coste de cada decisión. Esto se ha conseguido a través del concepto de beneficio de la solución que ayuda al usuario a tomar decisiones, las cuales incluso cuando no son las adecuadas aseguran la mínima pérdida. Esto permite al decisor obtener el máximo beneficio posible.

Para validar la metodología desarrollada, se aplica el modelo propuesto a cuatro bases de datos reales de ámbito médico, para así contrastar su efectividad. Los resultados obtenidos son comparados con los obtenidos por métodos conocidos como: J4.8 (la implementación en Weka del conocido C4.5), RBF-Networks, k-NN (k=3, 9)y RBC convencional. Los resultados se evaluaron en precisión, sensibilidad y especificidad. Los experimentos mostraron que nuestro modelo consigue el mismo nivel de precisión que los otros; puesto que no hay diferencias significativas en este sentido. Si se obtuvieron buenos resultados en sensibilidad, lo que supone, en el contexto en que nos movemos un bajo coste.

Capítulo 6

Aplicaciones a Problemas Reales

En este capítulo se aplica la metodología desarrollada en esta memoria a dos situaciones reales. En una de ellas, a partir de la Información de Riesgo y la probabilidad de una solución, se estudia la relación entre la flexibilidad y la estrategia de operaciones en una muestra real de empresas consultoras de ingeniería en España [11]. La otra aplicación es una adaptación del concepto de riesgo al diseño de nuevos productos [23].

6.1. Introducción

A lo largo de este trabajo se ha desarrollado una nueva metodología que mejora la etapa de recuperación de un sistema de RBC. Dicha metodología ha sido probada experimentalmente frente a otros métodos, obteniendo buenos resultados. Adicionalmente se han desarrollado dos aplicaciones prácticas en problemas con riesgo: diseño de nuevos productos y toma de decisiones en la empresa. En los dos problemas se adaptó la metodología desarrollada según las necesidades de cada uno. En este capítulo se estudiaran de forma detallada estas dos aplicaciones, viendo el proceso descrito de adaptación, y el resultado obtenido.

El resto del capítulo queda organizado de la siguiente manera: en la Sección 2, a través de los conceptos de probabilidad inicial de una solución y probabilidad condicionada, se desarrolla un sistema para estudiar las relaciones entre la estrategia de operaciones y la flexibilidad de una empresa. Más tarde, en la Sección 3, se presenta una aplicación del modelo de recuperación desarrollado en el capítulo 3 al problema del diseño de nuevos productos, el objetivo de esta aplicación era diseñar

un nuevo modelo de audífono con la empresa Beltone. Finalmente, en la Sección 4 se exponen las conclusiones del capítulo junto a un resumen del mismo.

6.2. La Información de Riesgo aplicada al estudio de la relación entre la Flexibilidad y la Estrategia de Operaciones

El estudio de la estrategia de operaciones y la flexibilidad ha tomado auge en las últimas décadas debido principalmente a los cambios constantes que sufre el entorno empresarial. Dichos cambios impulsan a las empresas al desarrollo de nuevas técnicas de toma de decisiones que les ayuden a adaptarse rápidamente. En este contexto, el saber cómo se relacionan la estrategia de operaciones y la flexibilidad, juega un papel muy importante, sobre todo a la hora de aumentar el rendimiento general de la compañía y por tanto su competitividad. Comprender la mejor forma de ajustar la estrategia de operaciones y la flexibilidad es crucial para disminuir la incertidumbre en el proceso de toma de decisiones de la empresa. Por este motivo la relación entre la estrategia de operaciones y la flexibilidad se ha convertido en el objeto de numerosos estudios científicos [143, 56].

Fundamentalmente, estas relaciones se han analizado a través del uso de diferentes métodos de análisis estadístico [9], aunque en los últimos años han aparecido estudios que se basan en distintas técnicas de Inteligencia Artificial (IA), como sistemas expertos [114, 152], RBC [104, 108], algoritmos genéticos [4], etc. Por tanto, resulta interesante analizar y estudiar la relación entre la estrategia de operaciones y la flexibilidad con técnicas de IA, que aporten un valor distinto y añadido en sus resultados, del que puedan aportar los obtenidos a través de la estadística tradicional. Este problema ha sido estudiado previamente por nosotros, como puede verse en [10], donde se construyó un sistema experto difuso y un asistente para la generación de reglas (ESROM), con el que se simulaban distintas situaciones con el objetivo de facilitar al gerente de la empresa la toma de decisiones.

En [11] se construyó un sistema RBC utilizando la metodología presentada en esta memoria, al que notaremos FP-RBC, que hace uso tanto de la probabilidad, como de la lógica difusa. Este sistema, utiliza la información que sobre una solución aportan la probabilidad inicial y la condicionada (definidas en el capítulo 5), y además,

se aprovecha de la flexibilidad con que la lógica difusa contribuye al problema. La elección de un sistema RBC para este tipo de problemas resulta acertada puesto que es una herramienta sencilla a la hora de ser puesta en práctica y que puede trabajar con problemas complejos y no estructurados. Además, su modelo de predicción se mantiene en un estado actualizado debido a que el caso base es revisado en tiempo real, que es un característica muy importante para una aplicación práctica.

En esta sección se verá una descripción completa del sistema que se desarrolló, los resultados obtenidos y la interpretación que de estos resultados hizo el experto.

6.2.1. Descripción del sistema

En esta sección se describe paso a paso como trabaja el sistema desarrollado, el cual hace uso, tanto de la lógica difusa como de la Teoría de la probabilidad para mejorar su rendimiento y aumentar su precisión. El FP-RBC, (cuyo objetivo es descartar los atributos del caso que no aportan información relevante en la decisión), parte de las probabilidades iniciales de cada solución, estas probabilidades son calculadas según Definición 5.1. Una vez calculadas, actualiza esta información, con la información que aporta cada valor concreto del atributo, utilizando para ello la Definición 5.2 de probabilidad condicionada. Ya calculadas estas probabilidades condicionadas y eliminado los atributos que no aportan suficiente información, se define el índice de información, concepto definido de forma paralela al de información de riesgo pero adaptado a este contexto. Este índice se introduce dentro de la medida de similitud usando un sistema de inferencia difuso que modificará los valores finales de dicha medida en función de los valores del índice de información. La salida del sistema nos dará aquella información que realmente aporte valor y eliminará la que no resulte interesante o no aporte verdadero valor a la hora de tomar una decisión.

El sistema consta de 7 pasos, veamos a continuación el proceso de forma detallada. Sea B una base de casos con k casos $C_1, \ldots, C_k$ y m soluciones diferentes $S_1, \ldots, S_m$. Cada caso contiene n atributos $C = (A_1, \ldots, A_n)$. Un caso nuevo llega al sistema $Caso\ Nuevo = (A_1 = a_1, \ldots, A_n = a_n)$, donde $a_1, \ldots, a_n$ son los valores concretos que toma cada atributo de *Caso Nuevo*

Paso 1. Cálculo de las probabilidades iniciales de cada solución

El primer paso es calcular la probabilidad inicial, $P_0(S)$, asociada a cada solución S almacenada en la memoria. Esta probabilidad se calcula con la información alma-

cenada en la base de casos y la Definición 5.1 que aparece en la Sección 5.2.1.1 del capítulo 5.

Paso 2. Selección de los atributos informativos a través de la probabilidad condicionada a cada solución

Una vez calculada la probabilidad inicial de la solución, ésta se actualiza en función de la información que aportan los valores de los atributos del caso nuevo mediante reiteradas aplicaciones de la definición de probabilidad condicionada (Definición 5.2) y del Teorema de Bayes, hasta obtener $P(S_i/Caso\ Nuevo)$. Para ello, introducimos los atributos uno a uno y calculamos las respectivas probabilidades. Si el resultado de ésta probabilidad es aproximadamente igual a cero, este atributo automáticamente es eliminado de la probabilidad. Se utiliza esta simplificación ya que en esta aplicación se interpreta la probabilidad como una información sobre el problema, por esta razón, si la probabilidad es cero, el atributo no contribuye a la información y se calcula la probabilidad de los siguientes, sin tener este atributo en cuenta. Vamos a ver este proceso en detalle.

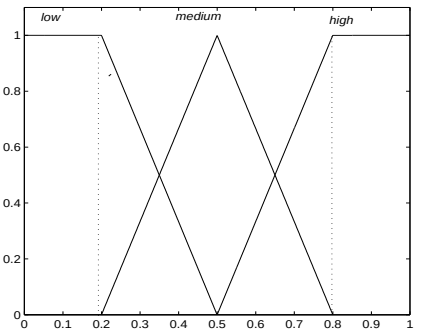
Comenzamos con el atributo 1 y calculamos $P_1(S_i) = P(S_i/A_1 = a_1)$. Una vez conocido el resultado, comprobamos si supera o no el *corte de información*; como corte de información se ha fijado 0.001. Si el resultado supera el *corte de información*, entonces se introduce el siguiente atributo y se calcula $P_2(S_i) = P(S_i/A_1 = a_1 \cap A_2 = a_2)$. Si no es así, lo eliminamos de la probabilidad y calculamos la siguiente sin tenerlo en cuenta $P_2(S_i) = P(S_i/A_2 = a_2)$ y así continúa el proceso hasta llegar al atributo n-*ésimo*, $P_n(S_i) = P(S_i/A_1 = a_1 \cap, \ldots, \cap A_n = a_n) = P_n(S_i/New\ Case)$. Cada P_i se calculará igual que se calculaba la probabilidad condicionada en el capítulo 5, en este cálculo, el número de atributos que intervengan quedará marcado por el *corte de información*.

Paso 3. Índice de Información

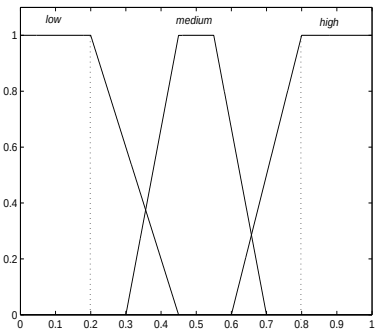
El *Índice de Información*, $I = (I_1, \ldots, I_n)$, será una medida sobre cuanto informa un atributo concreto, es un vector en el que cada componente I_i guarda el número de veces que la probabilidad condicionada a ese atributo supera el corte de información.

Paso 4. Cálculo de la similitud local

Una vez calculado el *Índice de Información*, I, se calcula la similitud entre atributos con las medidas dadas en Ecuación 3.1 y Ecuación 3.2 dependiendo de si el atributo es discreto o continuo.



(a) Función de pertenencia para la similitud



(b) Función de pertenencia para el Índice de Información

Figura 6.1: Funciones de pertenencia

Paso 5. Asignación de las etiquetas difusas

En este paso se asignan las etiquetas difusas a la similitud local y al índice de información (I), para ello, se han utilizado las siguientes funciones de pertenencia (Figura 6.1). No supone ninguna restricción utilizar cualquier otra que se ajuste a los datos.

Paso 6. Modificar la similitud local

Se introduce el Índice de Información (I_i) en la medida de similitud, siguiendo el mismo procedimiento que en capítulos anteriores. Para ello, se modifica la fórmula de la similitud haciéndola depender del Índice de Información de cada atributo (I_i) y de la similitud local ($sim(x_i^{Mem}, x_i^{Nue})$). Permitiéndonos así obtener la similitud modificada, $sim^{I_i}(x_i^{Mem}, x_i^{Nue})$:

$$sim^{I_i}(x_i^{Mem}, x_i^{Nue}) = f(I_i, sim(x_i^{Mem}, x_i^{Nue})) \tag{6.1}$$

La función $f(\cdot)$ es obtenida a través de un sistema de inferencia difuso. El sistema de inferencia difuso usado en la Ecuación 6.1 es también una aplicación del TSK model [163]. Este sistema para el i-*ésimo* atributo contiene las siguientes 9 reglas:

Regla 1. Si I_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces
v_i=$sim(x_i^{Mem}, x_i^{Nue}) + 6 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 2. Si I_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces v_i=$sim(x_i^{Mem}, x_i^{Nue}) + 5 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 3. Si I_i es *alto* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces v_i=$sim(x_i^{Mem}, x_i^{Nue}) + 2 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 4. Si I_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces v_i=$sim(x_i^{Mem}, x_i^{Nue}) - 0.2 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 5. Si I_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces v_i=$sim(x_i^{Mem}, x_i^{Nue}) - 0.3 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 6. Si I_i es *medio* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces v_i=$sim(x_i^{Mem}, x_i^{Nue}) - 0.4 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 7. Si I_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *alto*, entonces v_i=$sim(x_i^{Mem}, x_i^{Nue}) - 0.3 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 8. Si I_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *medio*, entonces v_i=$sim(x_i^{Mem}, x_i^{Nue}) - 0.4 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Regla 9. Si I_i es *bajo* y $sim(x_i^{Mem}, x_i^{Nue})$ es *bajo*, entonces v_i=$sim(x_i^{Mem}, x_i^{Nue}) - 0.5 \cdot sim(x_i^{Mem}, x_i^{Nue}))$

Como puede verse en las reglas arriba descritas, el sistema modifica a la medida de similitud en función del índice de información y de la similitud local. Veamos en detalle el por qué del diseño de algunas de ellas. Por ejemplo en la *Regla 1*, el índice de información toma el valor *alto*, esto significa que el número de veces que este atributo con ese valor superó el *corte de información* es alto, esto indica que se trata de un atributo que aporta información, si además es muy similar al del caso considerado, resultará interesante que influya en la recuperación más que cualquier otro que aporte menos información. Por el contrario, las *Reglas 7-9* tienen un valor

del índice de información *bajo*, esto lo interpretamos como que ese atributo apenas aporta información al caso por lo que hacemos que a la hora de recuperar no tenga mucha importancia; la importancia como puede verse en las reglas la bajamos en función de lo similar que sea al caso considerado.

Paso 7. Cálculo de la similitud modificada

El cálculo de la similitud o equivalentemente la inferencia del sistema, se hace en dos pasos. Primero, se calcula la fuerza de disparo de cada regla, y luego, se hace la media ponderada de las salidas de todas ellas. El procedimiento es el mismo que capítulos anteriores, una descripción detallada puede encontrarse en la Sección 3.2 del capítulo 3.

6.2.2. Descripción del problema: Caso Práctico

El objetivo de esta aplicación es establecer relaciones entre la flexibilidad y la estrategia de operaciones para el soporte en la toma de decisiones en una empresa. Para tal fin, se aplica la herramienta desarrollada a una muestra de 71 empresas de consultoría en España y se extraen conclusiones sobre los resultados obtenidos.

6.2.2.1. Los datos

El experto proporcionó datos recogidos acerca de la estrategia de operaciones, el nivel de flexibilidad y el rendimiento de una muestra de 71 empresas de consultoría de ingeniería en España. Un cuestionario fue la técnica utilizada para obtener los datos para el estudio. El cuestionario consta de 122 preguntas, divididas en 3 grupos principales, a su vez divididos en distintos bloques de preguntas cada uno. El primer grupo hace referencia a la estrategia de operaciones y está dividido en 9 bloques, el segundo grupo corresponde a la flexibilidad y está divido en 7, hay un tercer grupo que no usaremos en esta aplicación que corresponde el rendimiento de la empresa. Ver el cuestionario completo en Apéndice B. Tabla 6.1 y Tabla 6.2 muestran las variables que se usaron en el estudio y los items con los que se corresponden en el cuestionario. Bajo la columna *Variable* aparecen los nombres de las variables y la columna *Bloque* indica los bloques del cuestionario en los que aparecen estas variables.

Tabla 6.1: Estrategia de operaciones

Variable	Bloque	Variable	Bloque
Distribución en planta	AI	Actividades de trastienda y cara al cliente	AVI
Orientación Push/Pull	AII	Recursos Humanos	AVII
Nivel de estandarización	AIII	Participación del cliente	AVIII
Abanico de servicios	AIV	Diseño y desarrollo de nuevos servicios	AIX
Uso de tecnologías de la información	AV		

Tabla 6.2: Flexibilidad

Variable	Block	Variable	Block
Expansión	B1-B6	Mercado	B17-B18
Distribución de la información	B7-B9	Servicios y productos	B19-B20
Routing	B10-B12	Procesos, programación y volumen	B21-B24
Equipment	B13-B16		

6.2.2.2. Ejecución de los experimentos

El sistema ha sido probado usando validación cruzada de 5 subconjuntos. En cada validación, como ya se ha explicado en otros capítulos, se toma la base de datos entera y se divide en 5 conjuntos excluyentes entre si y con la misma distribución. Cada subconjunto se usa una vez como test, para ver los resultados que se obtienen al probarlo contra el conjunto que resulta de unir los otros cuatro restantes subconjuntos. Se hicieron 10 validaciones completas, lo que supone un total de 50 repeticiones de cada experimento para tener datos suficientes y poder así contrastar los resultados con t-test . Evaluamos los resultados obtenidos según precisión, puesto que el objetivo de esta aplicación es tener la evidencia suficiente para apoyar y justificar las conclusiones obtenidas.

El FP-RBC se implementó en MATLAB 7.1. y para contrastar la eficiencia de este método, los resultados obtenidos se compararon con los obtenidos por los siguientes modelos: J4.8, Naïve Bayes, Logistic (de los que se ha utilizado la versión disponible en WEKA [182]) y el estándar RBC implementado a propósito en MATLAB 7.1.

6.2.3. Resultado del estudio y discusión

Finalmente, se verán los resultados obtenidos y las conclusiones extraídas por el experto.

6.2.3.1. Resultados experimentales

En este apartado se muestran los resultados del caso de estudio. La Tabla 6.3 muestra las dimensiones de las operaciones estratégicas que superan el *Índice de Información* (IR). La Tabla 6.4 muestra un resumen de los resultados obtenidos en precisión para cada dimensión de flexibilidad.

Tabla 6.3: Grupo de atributos seleccionado en la dimensión Estrategia de Operaciones

VC	FB1	FB2	FB3	FB4	FB5	FB6	FB7
1	AI,AIV AIX	AI,AIX	AII,AIII AIV,AVIII	AII,AIII AIV	AI,AII AV	AIX	AI
2	AI,AIV AIX	AIX	AIV,AVIII	AII,AIII AIV,AIX	AI,AII	AIX	AI,AIX
3	AI,AIV AIX	AIX	AIV,AVIII	AI,AII,AIII AIV,AV,AVI	AI,AII AIII	AIX	AI,AIX
4	AI,AIV AIX	AIX	AI,AII,AIII AIV,AVIII	AII,AIII AIV	AI,AII	AIX	AI,AIX
5	AI,AIV AIX	AV,AIX	AIV,AVIII	AII,AIII	AI,AII	AIX	AI,AII AIX
6	AI,AIV AIX	AI,AIX	AIV,AVIII	AII,AIII,AIV AVI,AVIII,AIX	AI,AII AVIII	AIX	AI,AVI
7	AI,AIV AIX	AIX	AIV,AVIII	AIII	AI,AII,AIII AVIII,AIX	AIX	AI
8	AI,AIV AIX	AV,AIX	AIV,AVIII	AI,AII,AIII AIV,AVI,AIX	AI,AII	AIX	AI,AIX
9	AI,AIV AIX	AIX	AIV,AVIII	AI,AII AIII,AIV	AI,AII	AIX	AI,AIX
10	AI,AIV AIX	AIX	AIV,AVIII	AIII	AI,AII AV	AIX	AI,AIX

Como se puede observar FP-CBR es el método que obtiene mayor precisión de entre todos los métodos y en todas las dimensiones. Aplicamos t-test para verificar que los resultados son estadísticamente significativos. Tabla 6.5 muestra estos resultados.

Tabla 6.4: Precisión media de cada clasificador

	J4.8	**Bayes**	**Logistic**	**RBC**	**FP-RBC**
FB1	0.8663	0.8894	0.8707	0.8584	0.8923
FB2	0.8083	0.7649	0.7074	0.7074	0.8233
FB3	0.7322	0.5786	0.7353	0.6707	0.7384
FB4	0.7077	0.7231	0.7123	0.7138	0.7467
FB5	0.6653	0.6908	0.6526	0.6944	0.7017
FB6	0.7483	0.8090	0.7282	0.8108	0.7449
FB7	0.8416	0.8283	0.8433	0.7216	0.8600

Tabla 6.5: Resultados de los t-test al contrastar FP-RBC con cada modelo

	J4.8-FPCBR	**Bayes-FPCBR**	**Log-FPCBR**	**CBR-FPCBR**
FB1	0.023	0.749	0.055	0.016
FB2	0.041	0.000	0.000	0.000
FB3	0.493	0.000	0.846	0.003
FB4	0.047	0.05	0.166	0.086
FB5	0.138	0.415	0.010	0.020
FB6	0.000	0.891	0.000	0.000
FB7	0.435	0.004	0.053	0.000

6.2.3.2. Discusión de los resultados

Por último, el conocimiento mostrado en la Tabla 6.3 de acuerdo a los elementos y dimensiones utilizadas en [2] se puede resumir en:

1. *AI(Distribución en planta), AIV(Abanico de servicios ofrecidos), AIX(Diseño y desarrollo de nuevos servicios) está directa y fuertemente influenciado por FB1(Expansión).* Las decisiones estratégicas con respecto a las modificaciones físicas, los cambios en el número de servicios y/o en el desarrollo de nuevos servicios están directamente relacionadas con la expansión. Esta dimensión aumenta la flexibilidad cuando el coste y el tiempo de ampliación de capacidad son moderados. Sin embargo, hay un límite en la expansión de capacidades. Una vez alcanzado ese límite, el outsourcing puede ser una opción para aumentar la flexibilidad sobre ese límite.

2. *AIX(Diseño y desarrollo de nuevos servicios) está directa y fuertemente influenciado por FB2(Distribución de la información).* El desarrollo de nuevos servicios implica una redefinición de la distribución de la información dentro

del área de operaciones de la compañía. La naturaleza de esta redefinición será directamente relacionada con el número de recursos y capacidades de los nuevos servicios compartidos con los ya existentes. Las compañías que compiten en servicios innovadores necesitan altos niveles de flexibilidad en la distribución de información.

3. *AIV(Abanico de servicios ofrecidos), AVIII(Participación del cliente) está directa y fuertemente influenciado por FB3(Routing).* Cuando la compañía aumenta el número de servicios ofrecidos así como la participación de los clientes en la prestación de servicios, el ajuste de máxima flexibilidad se relaciona directamente con las decisiones estratégicas. Cuando se ofrecen nuevos servicios, algunas actividades son comunes a servicios antiguos. La necesidad de crear valor en los nuevos servicios aumenta el número de actividades nuevas que deben ser puestas en marcha, y aparecen nuevos ajustes en las actividades que deben ser desarrolladas. La participación del cliente implica un grado de variabilidad en la forma en la que los servicios son suministrados. En ocasiones, se necesita ofrecer distintos tipos de recursos a los clientes para su autoservicio. Esto incrementa la necesidad de una mayor flexibilidad en los ajustes.

4. *AIII(Nivel de estandarización) está directa y fuertemente influenciado por FB4(Equipment).* La dimensión de la estrategia que se corresponde con el grado de estandarización está directamente relacionada con la personalización del equipamiento. Un alto grado de estandarización requiere un diseño de servicio basado principalmente en el uso general de los recursos de equipamiento. Por otro lado, la personalización implica el intercambio de información entre el cliente y el personal para adaptar el servicio a las necesidades y deseos del cliente.

5. *AI(Distribución en planta), AII(Orientación Push/Pull) está directa y fuertemente influenciado por FB5(Mercado).* La distribución en planta y la orientación push/pull están directamente relacionadas con la dimensión de flexibilidad referente al mercado. La adaptación a los cambios del mercado requiere la correcta combinación de la distribución en planta así como de una correcta orientación Push/Pull para minimizar todo lo posible las demoras y los servicios no atendidos.

6. *AIX(Diseño y desarrollo de nuevos servicios) está directa y fuertemente influenciado por FB6(Servicios y productos).* El diseño y desarrollo de nuevos servicios estratégicos y la dimensión de flexibilidad referente a los productos y servicios están profundamente interrelacionadas. Las decisiones estratégicas para el suministro de nuevos servicios implican altos niveles de producción y flexibilidad para minimizar los costes de intercambio desde los viejos a los nuevos servicios, incrementándose los valores de salida.

7. *AI(Distribución en planta) está directa y fuertemente influenciado por FB7 (Procesos, programación y volumen).* Finalmente la dimensión de flexibilidad referente al proceso, programación y volumen está directamente relacionada con la distribución en planta. Diferentes configuraciones permiten operar a diferentes niveles de valores de salida de acuerdo con los diferentes arreglos en las dimensiones de procesos, programación y volumen.

6.3. La Información de Riesgo aplicada al desarrollo de nuevos productos

La gran competitividad a la que está sometido el mercado actual y los rápidos avances tecnológicos, obligan constantemente a las empresas a desarrollar nuevos productos. Por lo que para no perder su lugar dentro del mercado les surge el problema de cómo mejorar tanto el coste como el tiempo de desarrollo y fabricación de un nuevo producto. Esto unido al hecho de que adaptar los productos a las nuevas tecnologías y a las cada vez mayores exigencias de los consumidores hacen que su desarrollo se haga más complejo y que su fabricación sea más difícil y costosa.

En este contexto, puede resultar interesante el desarrollo de un sistema para diseño de nuevos productos. Los métodos más extendidos para tal fin son técnicas estadísticas, normalmente no se usan técnicas basadas en Inteligencia Artificial. Sin embargo, cuando surge la necesidad del nuevo producto, gran parte de la responsabilidad recae sobre la experiencia del experto humano. Es por esto que un sistema RBC puede ser una herramienta apropiada para este tipo de problemas. En la literatura existen algunas aplicaciones de RBC en este campo [120, 158, 185, 186] con buenos resultados. Al incluir la Información de Riesgo en la etapa de recuperación,

se obtuvieron buenos resultados lo que nos ha llevado a adaptar esta técnica como complemento en el diseño de nuevos productos.

Por esto en [23] se diseñó un modelo para diseño de nuevos productos, modelo que se puso en práctica con Beltone, multinacional que se dedica a la fabricación de audífonos. Para construir este modelo se definió un concepto al que se llamó *Adaptación de una Característica*, que era una adaptación de la Información de Riesgo al problema de diseñar nuevos productos, concretamente audífonos. En lugar de medir si una solución es o no adecuada para resolver un problema determinado, lo que hace es medir si el atributo será o no adecuado para para formar parte del nuevo diseño en función de las características del nuevo diseño. Esto puede ahorrar tiempo y entrenamiento del personal encargado de diseñar y fabricar nuevos productos.

En esta sección veremos primero cómo se adaptó nuestro modelo de recuperación basado en riesgo para este problema concreto. A continuación veremos cómo se comporta frente a un problema real aplicándolo al diseño de un nuevo modelo de audífono.

6.3.1. Un sistema RBC para desarrollo de nuevos productos

Veamos, ahora de forma detallada, como se adaptó el método presentado en el Capítulo 3 al diseño y desarrollo de nuevos productos. La base de casos se corresponderá con el portafolio de productos de la compañía, por tanto, un caso será un producto y sus atributos cada una de las características de dicho producto. Cuando un nuevo producto vaya a ser desarrollado, el producto o productos más parecidos de la base de casos serán recuperados. Para minimizar el coste y el tiempo de fabricación, se utiliza como base del nuevo producto el que menos modificaciones necesite de entre todos los productos del catálogo. Por tanto, elegir el caso más adecuado para resolver el problema, se traduce en como elegir el caso que minimice el tiempo y el coste de producción y que además evite que surjan incompatibilidades entre las diferentes características del nuevo producto. Para ello, el concepto de información de riesgo se adapta a la situación y pasa a llamarse ahora *Adaptación de una Característica.*

6.3.1.1. Definición de la Adaptación de una Característica

En esta sección se define la Adaptación de una Característica, de forma paralela a como se definió el concepto de Información de Riesgo; y se construye el sistema, teniendo en cuenta los aspectos importantes del problema concreto al que nos enfrentamos.

Supongamos que $P = \{P_1, \ldots, P_q\}$ es el portafolio de productos de la empresa y $C = \{c_i\}$, $i = 1, \ldots, n$ el conjunto de todas las características o propiedades que presentan dichos productos. $V_c = \{v_{c_1}, \ldots, v_{c_n}\}$, el conjunto de valores que puede tomar una característica concreta del producto; $\Theta = \{(Coste/U, Hora/U)_i\}$, $i = 1, \ldots, n$, el conjunto de vectores formados por los valores $Coste/U$ que es el precio que cuesta fabricar para un producto una característica concreta y se mide en euros y $Hora/U$ que es el tiempo que se tarda en fabricar para un producto una característica y se mide en horas. Ω es el conjunto de etiquetas lingüísticas definidas para el tipo de adaptación de cada característica al problema. En nuestro modelo hemos supuesto $\Omega = \{bajo, medio, alto\}$.

Definición 6.1. La adaptación de una característica del producto AC, se define $\forall\ c \in C$ como la aplicación $AC : V_c \times \Theta \rightarrow \Omega$ que mide la influencia positiva de esa característica sobre el nuevo producto para cada par de valores $(v_c \in V_c, (Coste/U, Hora/U) \in \Theta)$.

Esta definición lo que indica es la influencia positiva que esa característica tiene en la fabricación o diseño del nuevo producto en función de los valores Coste/U y Hora/U, es decir, da información de si es o no adecuado en función de su coste y tiempo de fabricación el introducir esa característica al nuevo producto.

Definición 6.2. Llamaremos, *Información de Adaptación*, al conjunto de aplicaciones $\{AC_i\}$, $i = 1, \ldots, n$, donde n es el número de características para cada producto.

Por tanto, la *Información de Adaptación*, no es más que el conjunto de las adaptaciones de una característica de un producto. Al ser el desarrollo de nuevos productos un problema que pertenece a la clase *Problemas Con Riesgo* resulta adecuado aplicar la *Información de Riesgo* para resolverlo.

6.3.1.2. Una nueva medida de similitud basada en la adaptación local

Para hacer uso de la *Información de Adaptación* se introduce en la medida de similitud, siguiendo el mismo procedimiento utilizado en el capítulo 3 para la *Información de Riesgo*. Para ello, se modifica la fórmula de la similitud global, ver Ecuación 6.2, haciéndola depender de Adaptación de la Característica (AC_i), del coste por unidad $Coste/U_i$ y de las horas por unidad $Hora/U_i$. Estos elementos solo intervienen en el sistema y no en el cálculo de la similitud.

$$Sim(P^{Mem}, P^{Nue}) = \frac{\sum_{i=1}^{n} f(AC_i, Coste/U_i, Hora/U_i)}{n} \tag{6.2}$$

donde n es el número de atributos o características del producto, P^{Mem} uno de los audífonos del catálogo y P^{Nue} el nuevo producto a desarrollar. La función $f(\cdot)$ es implícitamente obtenida con el sistema de inferencia difuso y es calculada como la media ponderada de las salidas de todas las reglas. El sistema de inferencia difuso usado en la Ecuación 6.2 es también una aplicación del modelo TSK [163]. Este sistema para el i-ésimo atributo contiene las siguientes 19 reglas:

Regla 1. Si AC_i es *alto* y $Coste/U_i$ es *alto* y $Hora/U_i$ es *alto*, entonces v_i=$2 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 2. Si AC_i es *alto* y $Coste/U_i$ es *alto* y $Hora/U_i$ es *medio*, entonces v_i=$2.25 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 3. Si AC_i es *alto* y $Coste/U_i$ es *alto* y $Hora/U_i$ es *bajo*, entonces v_i=$2.5 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 4. Si AC_i es *alto* y $Coste/U_i$ es *medio* y $Hora/U_i$ es *alto*, entonces v_i=$2.25 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 5. Si AC_i es *alto* y $Coste/U_i$ es *medio* y $Hora/U_i$ es *medio*, entonces v_i=$2.5 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 6. Si AC_i es *alto* y $Coste/U_i$ es *medio* y $Hora/U_i$ es *alto*, entonces v_i=$2.75 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 7. Si AC_i es *alto* y $Coste/U_i$ es *bajo* y $Hora/U_i$ es *alto*, entonces v_i=$2.75 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 8. Si AC_i es *alto* y $Coste/U_i$ es *bajo* y $Hora/U_i$ es *medio*, entonces v_i=$2.75 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 9. Si AC_i es *alto* y $Coste/U_i$ es *bajo* y $Hora/U_i$ es *bajo*, entonces v_i=$3 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 10. Si AC_i es *medio*, entonces $v_i = sim(A_i^{Mem}, A_i^{New})$

Regla 11. Si AC_i es *bajo* y $Coste/U_i$ es *alto* y $Hora/U_i$ es *alto*, entonces v_i=$0.2 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 12. Si AC_i es *bajo* y $Coste/U_i$ es *alto* y $Hora/U_i$ es *medio*, entonces v_i=$0.25 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 13. Si AC_i es *bajo* y $Coste/U_i$ es *alto* y $Hora/U_i$ es *bajo*, entonces v_i=$0.25 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 14. Si AC_i es *bajo* y $Coste/U_i$ es *medio* y $Hora/U_i$ es *alto*, entonces v_i=$0.3 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 15. Si AC_i es *bajo* y $Coste/U_i$ es *medio* y $Hora/U_i$ es *medio*, entonces v_i=$0.3 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 16. Si AC_i es *bajo* y $Coste/U_i$ es *medio* y $Hora/U_i$ es *bajo*, entonces v_i=$0.3 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 17. Si AC_i es *bajo* y $Coste/U_i$ es *bajo* y $Hora/U_i$ es *alto*, entonces v_i=$0.35 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 18. Si AC_i es *bajo* y $Coste/U_i$ es *bajo* y $Hora/U_i$ es *medio*, entonces v_i=$0.35 \cdot sim(A_i^{Mem}, A_i^{New})$

Regla 19. Si AC_i es *bajo* y $Coste/U_i$ es *bajo* y $Hora/U_i$ es *bajo*, entonces v_i=$0.5 \cdot sim(A_i^{Mem}, A_i^{New})$

En estas reglas, los términos lingüísticos, (*bajo, medio, alto*) están definidos sobre AC_i, $Coste/U_i$ y $Hora/U_i$, respectivamente; v_i $i = 1, \ldots, 19$ es la medida de similitud ajustada según la condición expuesta en la premisa de la i-ésima regla.

Las reglas se diseñaron con el objetivo de aumentar o disminuir la influencia de la similitud local sobre la similitud total, dependiendo del valor de la *Adaptación de la Característica* y del coste y tiempo de fabricación por unidad de esa característica. Por ejemplo, en la *Regla 9* se aumenta la similitud tres unidades, la razón es que siendo la adaptación de esa característica alta, (luego será sencillo adaptar el producto); tiene un coste y tiempo de fabricación asociados bajo. Ésta realmente es la situación perfecta, estamos ante una característica fácilmente adaptable y que además supone poco coste y requiere poco tiempo de fabricación. Por otro lado, la *Regla 10* es neutral, hemos considerado que si la adaptación es media, entonces no tiene importancia suficiente dentro del proceso de desarrollo para ser tenida en cuenta.

El cálculo de la similitud o lo que es equivalente la inferencia del sistema, se hace en dos pasos. Primero se calcula la fuerza de disparo de cada regla y luego la media ponderada de las salidas de todas ellas. El procedimiento está explicado de forma detallada en la Sección 3.2 del Capítulo 3.

6.3.2. Caso Práctico: Aplicación real para diseñar nuevos modelos de audífonos

Los retos y problemas que conlleva el diseño de un nuevo producto, son causados no sólo por el diseño en sí, sino también por los requerimentos del mercado. El corto

ciclo de vida que tienen los productos, junto con el reducido tiempo que las empresas tienen para su comercialización, desemboca en una alta competitividad que crea la necesidad de que los productos sean de alta calidad, alta funcionalidad y además precio bajo. Por estas razones, y con el objetivo de minimizar gastos, muchas empresas tienen que eliminar recursos, que se traduce en despedir miembros del equipo de diseño, reducir el tiempo de entrenamiento del personal, etc. Esto hace que la tarea de crear nuevos productos sea aun más difícil. Una mezcla de todos estos problemas son los que actualmente preocupan a la dirección de la empresa Beltone. Beltone es una compañía multinacional que se dedica a fabricar audífonos. Se contactó con su sede en España. Esta sede es un laboratorio donde se fabrican audífonos y se diseñan nuevos con el objetivo de captar nuevos clientes y aumentar la competitividad en el mercado. El proceso de fabricación de un audífono es un proceso delicado que requiere de personal cualificado. Los clientes de este laboratorio normalmente exigen nuevos diseños que puedan adaptarse a las tendencias del mercado y satisfacer sus expectativas a tiempo. Teniendo todo esto en cuenta, se le propuso como solución un modelo RBC ya que es adecuado para ayudar en la tarea de crear nuevos productos, puesto que el conocimiento almacenado del que se dispone puede utilizarse para adaptar los antiguos productos a los nuevos. Esto supone que es necesario menos personal y, que además éste necesitará de menos entrenamiento. Por tanto, la empresa mejorara en tiempo de fabricación y costes. El plazo de entrega para esta empresa es muy importante puesto que mantener los clientes depende en gran parte de lo rápido que sean capaces de servirles el producto. Veamos de forma detallada, el modelo propuesto.

- *Representación de los casos*

La organización de la memoria de casos se hizo con un experto de Beltone. La empresa tiene un portafolio de 350 modelos. El experto nos ayudó a definir un caso, C, como un conjunto de atributos $C = (A_1, \ldots, A_n)$. Cada atributo fue definido como un vector de dos dimensiones A= *(P, N)*, donde P (la característica) será una de las propiedades que tenga el producto y N serán los valores de coste y tiempo asociados a esa característica $N = (Coste/U_i, Hora/U_i)$, donde Coste/U se mide en euros y es el coste de cada propiedad P y Horas/U se mide en horas y es el tiempo de trabajo que necesita esa propiedad. Según el experto que nos ayudó a organizar la base de casos, se definieron 13 características como las más representantes: Control de Volumen (CV), Push Button (PB), Trimmers (TR), Hilo Extractor (HE), Switch (S), Filtro Anticerumen (FA), Punto de Referencia (PR), Bluetooth (B), Bobina

Tabla 6.6: Atributos del Nuevo Producto *I*

Atributos	**CV**	**PB**	**TR**	**HE**	**S**	**FA**	**PR**	**B**
Valores	?	?	No	No	No	Si	No	No

Tabla 6.7: (b) Atributos del Nuevo Producto *II*

Atributos	**BT**	**BD**	**MR**	**AE**	**TP**	**C/U**	**H/U**	**TF**
Valores	Si	Si	No	No	S312	650	1.5	3

Telefónica (BT), Bidireccionalidad (BD), Micrófono Remoto (MR), Auricular Externo (AE) y Tamaño de la Pila (TP).

- *Descripción del Problema*

Cuando al laboratorio le llega la orden de diseñar y fabricar un nuevo producto, les resultaría muy útil, en esta situación, disponer de un sistema que detectara cuál es el antiguo producto, que tomado como base del nuevo diseño, menos le cuesta en precio y horas trabajadas a la empresa. Veamos ahora qué ocurre si se utiliza el modelo propuesto para este fin. Lo aplicamos a un caso real, la construcción de un audífono con un circuito híbrido, que hace que no sea necesario que el modelo lleve ni control de volumen (propiedad 1) ni Push button (propiedad 2) puesto que el modelo híbrido lo que llevará incorporado será un sensor de ambiente que hará el papel de estas dos propiedades. El beneficio para el cliente consiste en que al reducir dos características en una, el tamaño se reduce y por tanto es más cómodo tenerlo instalado en su canal auditivo. Finalmente, y tras conocer todas las exigencias del cliente, el nuevo modelo que queremos construir es el siguiente: la Tabla 6.6 y la Tabla 6.7 muestran los valores de las características del nuevo producto. Ahora el jefe del equipo de desarrollo tiene que decidir cómo hacer este nuevo modelo con el menor esfuerzo, para producir el ahorro de costes y personal.

- *Recuperación del Caso*

Veamos ahora en detalle cómo recuperamos los casos más adecuados para la fabricación del nuevo producto. Primero calculamos la similitud entre los casos en memoria y el nuevo caso, *Modelo Nuevo*. Hemos escogido como medida de similitud la Heterogeneous Euclidean-Overlap Metric (HEOM) [181]. Esta medida devuelve la distancia entre dos casos C^{Mem} y C^{Nue} y se define como sigue:

$$Sim(C^{Mem}, C^{New}) = 1 - \sqrt{\frac{\sum_{i=1}^{n} sim_i^2(A_i^M, A_i^N)}{n}} \tag{6.3}$$

donde A_i^M y A_i^N son el i-ésimo atributo de C^{Mem} y C^{Nue} respectivamente y n es el número de atributos. La función $sim_i(A_i^M, A_i^N)$ usa una de las dos funciones definidas en Ecuación 6.4 dependiendo de si el atributo es nominal o numérico.

$$sim_i(A_i^M, A_i^N) = \begin{cases} 1 & si \;\; A_i^M \; o \; A_i^N \text{ es no conocido} \\ overlap(A_i^M, A_i^N) & si \;\; \text{i-ésimo atributo es nominal} \\ rn - diff_i(A_i^M, A_i^N) & si \;\; \text{i-ésimo atributo es numérico} \end{cases} \tag{6.4}$$

Los valores de los atributos no conocidos serán 1 por defecto, (que se corresponde con la máxima distancia). La función *overlap* y la función *rn–diff* se definen:

$$overlap(A_i^M, A_i^N) = \begin{cases} 0 & if \quad A_i^M = A_i^N \\ 1 & if \quad A_i^M \neq A_i^N \end{cases} \tag{6.5}$$

$$\text{rn–diff}_i(A_i^M, A_i^N) = \frac{|A_i^M - A_i^N|}{range_i} \tag{6.6}$$

El valor $range_i$ se usa para normalizar los atributos y se calcula como $range_i = max_i - min_i$, donde max_i y min_i son respectivamente, el valor máximo y mínimo observados en el conjunto de pruebas para el i-*ésimo* atributo.

Tabla 6.8 y Tabla 6.9 muestran los modelos recuperados, en ellas también puede verse el valor de sus atributos. Por otro lado, la Tabla 6.10 muestra los valores de la similitud global usando Ecuación 6.3. Como puede verse, los casos más similares son *Identity(IDT35)* y *Reach 35VC*. A primera vista ellos parecen los más adecuados para fabricar el Nuevo Modelo, ya que tienen las dos propiedades que buscamos (Control de Volumen Push Button), sin embargo no lo son. El experto nos dijo que si tomábamos como base un modelo con Control de Volumen la construcción del nuevo producto resultaría más cara que si elegimos un modelo sin esta característica. CV es una de las características más caras y más difíciles de fabricar. Veamos

Tabla 6.8: Casos recuperados *I*

	CV	**PB**	**TR**	**HE**	**S**	**FA**	**PR**	**B**
Reach RCH35	No	Si	No	No	No	Si	No	No
Identity(IDT35)	Si	Si	No	No	No	Si	No	No
Reach 35VC	Si	Si	No	No	No	Si	No	No
Identity (IDT45)	Si	No	No	No	No	Si	No	No

Tabla 6.9: Casos recuperados *II*

	BT	**BD**	**MR**	**AE**	**TP**	**C/U**	**H/U**	**TF**
Reach RCH35	No	Si	No	No	S13	625	1.75	3
Identity(IDT35)	Si	Si	No	No	S312	650	1.5	3.5
Reach 35VC	Si	Si	No	No	S312	600	1.5	3.25
Identity (IDT45)	Si	No	No	No	S312	650	1.8	3

Tabla 6.10: Similitud entre los casos en memoria y el Nuevo Modelo

	Similitud
Reach RCH35	0.8619
Identity(IDT35)	**0.8916**
Reach 35VC	**0.8915**
Identity (IDT45)	0.8749

ahora qué ocurre si utilizamos para seleccionar el modelo base la Información de Adaptación. Para ello, primero la asignamos de forma independiente a cada característica. Veamos a modo de ejemplo cómo sería la asignación de un caso, elegimos el *Reach RCH35*.

- Primer atributo: **Control de Volumen**
  AC_1 *(CV = No y* $Coste/U_1 = 60€$ *y* $Horas/U_1 = 50min$*) = Alto.* Porque CV es una característica cara y difícil y resultará mejor empezar el Modelo Híbrido sin ella, por tanto, el atributo *CV=No* tiene una fuerte influencia positiva en el Nuevo Producto.

- Segundo atributo: **Push Button**
  AC_2 (*PB=Si* y $Coste/U_2 = 30€$ y $Horas/U_2 = 10min$)= *Alto.* En este caso, el valor se debe a que se trata de una característica barata y además útil para el modelo híbrido. Por tanto *PB=Si* tiene una fuerte influencia positiva en el Nuevo Producto.

- Tercer atributo: **Trimmers**
  AC_3(*TR=Si* y $Coste/U_3 = 20€$ y $Horas/U_3 = 10min$)= *Medio.* Porque no toma parte en el proceso de fabricación.

El procedimiento es igual para los otros modelos y el resto de atributos. La Tabla 6.11 y la Tabla 6.12 muestra los valores de la adaptación para la adaptabilidad de cada atributo en cada modelo. En estas tablas, B es bajo y A es alto. En estas tablas no aparecen todos los resultados, ya que en la práctica solo los valores críticos son interesantes en el proceso, igual que ocurría con la Información de Riesgo. Por valores críticos entendemos aquellos que toman para la adaptación valor bajo o alto.

Ahora asignamos las etiquetas lingüísticas para coste por unidad y para hora por unidad, para así poder hacer inferencia en el sistema. En este caso no se asignan etiquetas lingüísticas a la adaptación puesto que ha sido directamente asignada con valor lingüístico. De no ser así, el proceso sería igual que para el coste y las horas por unidad. La Figura 6.2 muestra las funciones de pertenencia de coste por unidad y de hora por unidad. Se ha escogido la misma en los dos casos.

Vemos ahora como se disparan las reglas, se escoge un caso de la base de casos, por ejemplo *Reach RCH35* y una característica/atributo de este caso, que será Control de Volumen CV, $Coste/U_1 = 60€$ y $Horas/U_1 = 0.30$. Las etiquetas lingüísticas

Tabla 6.11: Tabla de asignación de la adaptación, sólo muestra los valores críticos *I*

	CV	**PB**	**TR**	**HE**	**S**	**FA**	**PR**	**B**
Reach RCH35	A	A	-	-	-	B	A	-
Identity(IDT35)	B	A	-	-	-	B	A	-
Reach 35VC	B	A	-	-	-	B	A	-
Identity (IDT45)	A	A	-	-	-	B	A	-

Tabla 6.12: Tabla de asignación de la adaptación, sólo muestra los valores críticos *II*

	BT	**BD**	**MR**	**AE**	**TP**	**C/U**	**H/U**	**TF**
Reach RCH35	B	-	-	-	-	-	-	-
Identity(IDT35)	B	B	-	-	-	-	-	-
Reach 35VC	-	B	B	-	B	-	-	-
Identity (IDT45)	B	B	-	-	-	-	-	-

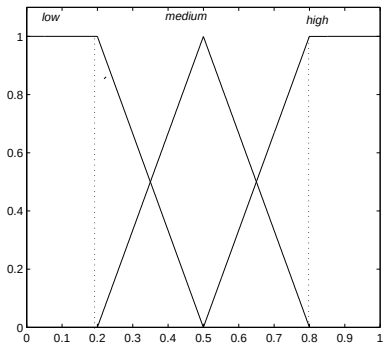


Figura 6.2: Función de pertenencia

para $Coste/U_1 = 60€$ y $Horas/U_1 = 50min$ pueden verse en Figura 6.2 y son *Alto* para los dos casos. La adaptación, nos dijo el experto, que es *alta* para ese atributo, por lo que se disparan: *Regla 1* y *Regla 5*.

Tabla 6.13: Similitud entre los casos en memoria y el Nuevo Modelo usando la adaptación

	Similitud
Reach RCH35	**0.9876**
Identity(IDT35)	0.7654
Reach 35VC	0.7432
Identity (IDT45)	0.9111

La fuerza de disparo de las reglas se calcula con Ecuación 3.6 para todas las reglas y Ecuación 3.7 para la 10. Finalmente, se obtienen los resultados mostrados en Tabla 6.13. Como puede verse, el modelo a recuperar es Reach RCH35, el cual es la mejor opción según el experto, puesto que no tiene CV y si tiene PB.

6.4. Resumen y conclusiones

Tras proponer un nuevo enfoque de recuperación en RBC para problemas que conllevan riesgo asociado, y ver su aportación en problemas reales, en este capítulo se han presentado dos aplicaciones. Las conclusiones de cada una se analizarán por separado.

La primera aplicación es un estudio para optimizar la gestión de recursos. Dicho estudio está basado en estudios previos realizados sobre la relación entre la estrategia de operaciones y la flexibilidad en una empresa. Adaptando el modelo de recuperación presentado en esta memoria, se ha desarrollado una aplicación que conduce a resultados interesantes, que son consistentes con los estudios previos [2,3]. Además, la aplicación, proporciona nuevas ideas que ayudan a comprender con mayor precisión, los efectos de la relación entre la estrategia de operaciones y la flexibilidad. Para futuras investigaciones, las técnicas basadas en Soft Computing, tales como los algoritmos genéticos o los sistemas difusos se incluirán para ajustar o aprender los parámetros de las reglas difusas, y las funciones de pertenencia. Estos parámetros son especialmente importantes en las etapas finales de desarrollo del sistema.

Por otro lado y debido a que la competitividad en el mercado actual crece constantemente, muchas compañías han reducido recursos para así seguir obteniendo beneficios. En esta situación resulta difícil el desarrollo de nuevos productos, que conlleva muchos gastos asociados. Por esto se ha aplicado el modelo de recuperación con riesgo presentado en este trabajo para desarrollar un prototipo. Este prototipo es un sistema RBC que actúa como soporte para el desarrollo de nuevos productos y que utiliza la experiencia del experto para desarrollar la tarea. Gracias al sistema, la experiencia del experto puede almacenarse en memoria y por tanto, puede ser usada en cualquier momento sin la necesidad de la presencia física del experto. Se muestra una aplicación real para diseñar audífonos. Se tiene previsto avanzar en la mejora del modelo.

Capítulo 7

Resumen, Conclusiones y Trabajos Futuros

En este capítulo final, se resumen los aspectos más destacables de este trabajo y el proceso que se ha realizado para llevar a cabo esta investigación. A continuación se exponen las conclusiones del trabajo, algunas ya han aparecido en los capítulos y otras son el resultado de toda esta tesis en su conjunto. Por último, se discutirán ciertas líneas de trabajo que pueden ser continuadas o que se han abierto durante esta investigación.

7.1. Resumen

En el *primer capítulo*, se expuso la motivación del trabajo, se detallaron los objetivos a conseguir y se esbozó la organización del mismo.

En el *segundo capítulo*, se analizó el concepto de Razonamiento Basado en Casos (RBC) desde una perspectiva general: estructura de un razonador basado en casos, historia, fundamentos, ciclo, funcionamiento, aplicaciones, e incluso, sus características más relevantes. El resto del capítulo se centra en el estudio de la etapa de recuperación, por ser precisamente en esta etapa donde se ha desarrollado esta investigación. Se analiza el concepto de similitud y se presentan aquellas medidas más relevantes. De forma análoga, se repasan las técnicas más significativas que actualmente se utilizan en recuperación, comenzando con métodos clásicos como el conocido principio del *Vecino más Cercano* (k-NN) o *Fish and Shrink* hasta llegar

a técnicas de recuperación más sofisticadas como las que usan algoritmos genéticos, redes neuronales o lógica difusa. Todo con el objetivo de situar la propuesta en un marco adecuado con un punto de partida actualizado.

En el *capítulo 3*, se discutieron las deficiencias que presentan algunos métodos actuales de asignación de similitud, cuando se enfrentan a problemas de la clase PCR. Y se realizó una propuesta original para contrarrestar algunas de ellas. Dicha propuesta consistió en la introducción de un nuevo concepto llamado *Información de Riesgo*. Este concepto se define de forma local, es decir, es asignado independientemente a cada atributo, tomando en cuenta el valor concreto del atributo y la solución del caso en memoria. Definir este nuevo concepto permite recuperar el caso más adecuado en lugar de simplemente el más similar. De acuerdo con lo establecido en el capítulo, por caso más adecuado se entiende aquel recuperado según el criterio de adecuación y no solo el de similitud. Además se realiza un estudio sobre las diferencias que existen entre los conceptos de riesgo y peso, para evitar cualquier tipo de confusión que pudiera surgir. Finalmente, para validar empíricamente las ventajas que aporta la Información de Riesgo en la etapa de recuperación, se compararon los resultados obtenidos utilizando este nuevo modelo con los resultados obtenidos por conocidos métodos de clasificación, en términos de precisión, y en cuatro bases de datos del UCI *machine learning repository*. Como pudo verse, los resultados demostraron que con un nivel de confianza del 95 % el sistema presentaba diferencias significativas en la precisión media obtenida, con respecto a los otros métodos, en tres de las bases que se usaron en los experimentos.

En el *cuarto capítulo*, una vez probadas las ventajas de considerar la función IRL en la recuperación del caso más adecuado, se analizan y estudian modelos, que permiten seguir haciendo uso de ella en ausencia total o parcial del experto. En el caso en que el experto proporcione parte de la información del problema, o se tenga información que provenga de muestras o ejemplos, se aplican conocidos modelos de Análisis Numérico que, por su buen comportamiento, justifican su elección. Estos modelos, son analizados para saber bajo qué condiciones, en función del tipo y cantidad de la información disponible, será adecuado utilizarlos. Por otro lado, en las situaciones en que no se tenga información alguna, se propuso un nuevo método, el cuál, haciendo uso de la información que proporcionan los casos almacenados en memoria, genera unos valores iniciales para el riesgo. Finalmente, estos métodos fueron probados bajo distintas hipótesis que simulaban situaciones reales, para analizar su comportamiento, y así poder extraer conclusiones de cuándo será ade-

cuado aplicarlos. Esto permite utilizar la IR en una mayor variedad de situaciones sacando, por tanto, el máximo provecho de ella.

En el *capítulo cinco*, se estudia y analiza la influencia de las consecuencias asociadas a cada decisión, pero ahora, dividiendo estas consecuencias en términos de ganancia y pérdida de cada solución. Para ello, a lo largo del capítulo se definen los conceptos de ganancia, pérdida y beneficio esperado de la solución considerada. Introduciendo estos conceptos en la medida de similitud global se consigue recuperar el caso con mayor beneficio, en el sentido en que minimiza el coste de la decisión. Para validar empíricamente el sistema, se compara en términos de precisión, sensibilidad y especificidad con conocidos métodos de clasificación. La comparación en sensibilidad se hace para ver cómo el sistema recupera en un número mayor de veces las soluciones cuyo coste es menor, es decir, recupera más casos correctos de la clase que mayor beneficio produce, de forma que cuando clasifica mal, lo hace en casos cuya solución es menos costosa. También se presentan estos resultados en términos de precisión para ver cómo, aunque el sistema incida en recuperar los casos de menor coste, no pierde en precisión. Es más, los resultados muestran que no existen diferencias significativas a un nivel de confianza del 95 % en términos de precisión, lo que se traduce en que el sistema sostiene el nivel de precisión y aumenta el de sensibilidad.

En el *capítulo seis*, se aplica la metodología presentada en esta memoria a problemas reales. Concretamente al diseño de nuevos productos y al estudio de la relación entre la flexibilidad y la estrategia de operaciones en las empresas. En el primer caso, se adapta la Información de Riesgo, y se convierte en Información de Adaptación, y esto se aplica a un problema real de la empresa Beltone. El sistema obtenido es aun un prototipo. La otra aplicación es un sistema, basado en la probabilidad de una solución y la Información de Riesgo, que estudia la relación entre la flexibilidad y la estrategia de operaciones en una muestra real de empresas consultoras de ingeniería en España, con el objetivo de facilitar la toma de decisiones los directivos de esas compañías.

7.2. Conclusiones

Desde el nacimiento del Razonamiento Basado en Casos, se han desarrollado, multitud de técnicas de recuperación que muestran, ante ciertos problemas, carencias en el enfoque que deberían solventarse para avanzar y obtener una conducta

más similar al razonamiento humano. Cuando un sistema de RBC resuelve un problema, generalmente lo que hace es buscar el más similar de entre todos los que tiene en memoria, y posteriormente, decide aplicar la solución de ese problema a la resolución del nuevo. Sin embargo, cuando nosotros hacemos ese razonamiento, también tenemos en cuenta qué consecuencias tendrá el resolver de forma errónea el problema. En muchas situaciones, estas consecuencias son fundamentales ya que aportan una información muy importante y en ocasiones determinante a la hora de decidir, sobre todo cuando se trata de resolver un problema que pertenece a la clase PCR. A lo largo de esta memoria se ha desarrollado una nueva metodología para recuperar en RBC con la que se da solución a este problema. Para exponer las conclusiones finales alcanzadas a lo largo del trabajo se responderá los interrogantes planteados al comienzo de esta memoria.

- ¿Qué criterios serán los adecuados para elegir el caso más conveniente y no solo el más similar?
- ¿Cómo se puede introducir esta información dentro del problema?
- ¿Cómo se comportará esta metodología comparada con otras ya existentes?
- ¿Qué ventajas aporta? ¿Cuáles serán sus debilidades?
- ¿Puede aplicarse a problemas reales?

Como criterios de selección del caso más adecuado se han utilizado dos conceptos definidos para tal fin como son la Información de Riesgo y el Beneficio esperado de la solución. Cuando el sistema se enfrena a un problema que pertenezca a la clase PCR, (es decir, cuando el sistema se enfrenta a problemas en los que aplicar una solución poco adecuada conlleva graves consecuencias) será necesario introducir esta información en el sistema a la hora de recuperar un caso, para que éste tenga en cuenta las consecuencias asociadas que conlleva aplicar la solución considerada al problema concreto. Como instrumento de medida de estas consecuencias nace el concepto de Información de Riesgo. Más tarde en un segundo paso se analizaron desde otra perspectiva estas consecuencias y se pudo ver como cada decisión que se toma lleva asociada una pérdida y una ganancia, por lo que se definen dos nuevas funciones encargadas de medir dichos conceptos. Estas funciones dan lugar al Beneficio esperado de cada solución. Concepto que informa del beneficio de que una cierta solución sea aplicada con éxito. Con estos dos nuevos conceptos se consigue un sistema algo más conservador, en el sentido en que evita recuperar aquellos casos

cuya solución pueda provocar graves consecuencias. De esta forma, cuando comete fallos, el sistema tiende a recuperar aquellas soluciones que aunque no dan solución correcta, al menos no provocan pérdidas importantes.

Una vez definidos estos conceptos se introdujeron dentro del problema para ser utilizados durante la etapa de recuperación. Para ello se utilizó un sistema de inferencia difuso. La integración de RBC con un sistema de inferencia difuso permite beneficiarse al sistema de las ventajas de ambos sistemas, como son:

- Su eficacia para tratar información imprecisa; esta ventaja es clave puesto que tanto el beneficio como para el riesgo suelen expresarse de forma imprecisa.
- Su capacidad para aportar a la medida información lineal y no lineal; esto le proporciona realismo ya que todo lo que influye sobre la similitud no tiene que ser de forma lineal.
- Facilita la recuperación de alternativas dentro de conjuntos grandes de casos; lo que permite complementar algunas limitaciones ya que con estas alternativas existe más variedad a la hora de recuperar, lo que permite al nuevo modelo tomar en la mayoría de casos uno de los que menos riesgo produce.

Desarrollada la metodología, se estudió su comportamiento, para ello, se comparó con conocidos métodos encontrados en la literatura de los que se tomó su implementación en WEKA y se utilizaron bases estándar tomadas del UCI *machine learning repository* [60]. Se hicieron varias pruebas obteniendo las siguientes conclusiones. Por un lado se probó como la Información de Riesgo frente a conocidos modelos como son: el J4.8 (la implementación en WEKA del conocido C4.5), RBF-Networks, ANFIS, k-NN ($k = 3, 9$) y RBC convencional. Obtuvo buenos resultados en precisión en problemas de la clase PCR. Los resultados demuestran cómo las diferencias en precisión eran estadísticamente significativas a un nivel de confianza del 95 %. Por tanto, esto demuestra que el nuevo concepto de riesgo mejora la recuperación en problemas de la clase *Problemas Con Riesgo* y guía a los pesos hacia una mejor recuperación. También se experimentó con el beneficio, para probar que cumplía su objetivo de minimizar el coste de la decisión. Se validó en términos de precisión, sensibilidad y especificidad también con conocidos métodos de clasificación: J4.8 , RBF-Networks, k-NN (k=*9*) y RBC convencional. En este experimento todas las bases que se utilizaron fueron de ámbito médico, para así probar cómo el modelo tiene la capacidad de evitar provocar peligro en la vida del paciente, en un número

alto de casos. Precisamente, la comparación en sensibilidad se hace para ver cómo el sistema recupera en un número mayor de veces las soluciones cuyo coste es menor, es decir, recupera más casos correctos de la clase que mayor beneficio produce. Lo que supone una ventaja, porque cuando el sistema clasifica mal, lo hace en casos cuya solución en menos costosa, lo que significa en este contexto que lo hace en situaciones en las que no pone en juego la vida del paciente. También se presentan estos resultados en términos de precisión para ver como, aunque el sistema incida en la recuperación de casos de menor coste, no pierde en precisión. Es más, los resultados muestran que no existen diferencias significativas a un nivel de confianza del 95 % en términos de precisión, lo que se traduce en que el sistema sostiene el nivel de precisión y aumenta, sin embargo, el de sensibilidad.

La metodología desarrollada se ha demostrado eficaz en problemas reales por lo que finalmente, para aportar valor añadido a esta investigación, se realizaron aplicaciones de esta metodología a problemas reales que estuviesen en la clase de problemas PCR. La primera de las aplicaciones fue un estudio para optimizar la gestión de recursos en la empresa. Para la realización de este estudio se adaptó el modelo de recuperación presentado en esta memoria. Obteniendo unos resultados interesantes ya que las nuevas ideas obtenidas ayudan a comprender con mayor precisión, los efectos de la relación entre la estrategia de operaciones y la flexibilidad. Lo que facilita la toma de decisiones en una empresa. Por otro lado y debido a la competitividad que actualmente existe en el mercado, muchas compañías han reducido recursos para así seguir obteniendo beneficios. En esta situación resulta difícil el desarrollo de nuevos productos, puesto que conlleva muchos gastos asociados. Por este motivo, dar soporte en el desarrollo de nuevos productos a estas compañías resulta una idea interesante y que puede reducir costes. Por estas razones, se ha aplicado el modelo de recuperación con riesgo presentado en este trabajo para desarrollar un prototipo, un sistema RBC que actúa como soporte para el desarrollo de nuevos productos y que utiliza la experiencia del experto para desarrollar la tarea. Gracias al sistema, la experiencia del experto puede almacenarse en memoria y por tanto, puede ser usada en cualquier momento sin la necesidad de que esté presente físicamente, lo que podría llevar a reducir coste de personal. Ahora el experto debe estar presente menos horas y el personal requiere de menos entrenamiento puesto que toda la información está guardada en el sistema. Finalmente con la ayuda de un experto de la empresa Beltone (empresa que se dedica al diseño y fabricación de audífonos) se desarrolla una aplicación real que ayuda a decidir cuáles son las características óptimas para poner en el mercado un nuevo audífono.

7.3. Trabajos futuros

Las líneas de investigación futuras se presentan agrupadas según la temática abordada por cada uno de los capítulos anteriores. No obstante, y como vía de estudio general se pretende seguir trabajando en esta línea, profundizado en el concepto de riesgo asociado a una solución y realizando aplicaciones prácticas en el campo de la medicina y la economía, por considerar que son campos de estudio, donde la metodología introducida tiene especial interés. Como ya se comentó, en problemas relacionados con estos campos, resulta provechoso tener en cuenta si la solución que se va a utilizar para resolverlo es o no adecuada.

En el capítulo 3 se presentó una técnica de recuperación, especialmente diseñada para aquellos problemas que pertenecía a la clase *Problemas Con Riesgo*. Esta técnica hacía uso de un nuevo concepto, llamado *Información de Riesgo* que permitía de forma independiente en cada atributo conocer las consecuencias de utilizar esa solución para resolver el problema. Su principal inconveniente era la dependencia con respecto al experto. En el siguiente capítulo se propusieron modelos que dieran solución a este problema, aún así se abre una línea importante de investigación en este sentido, la de aprender métodos de asignación de riesgo de forma automática en base a ejemplos, para así conseguir una mejor solución al problema de la asignación, válida para todo tipo de problemas y que además, nos aporte aún mas información acerca del problema a resolver. Actualmente se está desarrollando un sistema de ajuste para el riesgo que se adapta según evoluciona el sistema y el contexto, ajustando parámetros ya definidos en el sistema como las funciones de pertenencia y metiendo uno libre que nos permita controlar los resultados. También se considera el hacer un modelo que tenga en cuenta distintas informaciones de riesgo, bien porque viniesen de distintos expertos o porque aportasen informaciones distintas al problema.

En el capítulo 5, se profundiza en el estudio de las consecuencias que conlleva tomar una decisión en términos de la ganancia y la pérdida. Utilizando estos conceptos se define el beneficio de una solución. En esta memoria se ha definido de forma conservadora, como la diferencia entre la pérdida y la ganancia media. El introducir esta función de beneficio en el proceso de recuperación abre otra línea importante de investigación: la generación de modelos que definan la función beneficio adaptándose a la personalidad del usuario. De forma que si alguien muy atrevido usara el sistema, este se comportarse de forma arriesgada, importándole menos perder si

las ganancias son grandes. Sin embargo si alguien conservador lo usara esta función minimizaría la pérdida aunque supusiera esto obtener una menor ganancia.

Otra línea de trabajo a seguir es el estudio y desarrollo de aplicaciones reales de esta metodología, hay campos como la medicina o la empresa, donde tendría gran trascendencia el desarrollo de aplicaciones basadas en esta metodología.

Apéndices

Apéndice A

Información de Riesgo: Ejemplo Práctico

En este apéndice, se explica en detalle como se aplica el modelo con riesgo al ejemplo presentado a modo de introducción en el capítulo 3. Este ejemplo sencillo servirá para ver fácilmente el funcionamiento del modelo. El primer paso será asignar la Información de Riesgo, para ello se utiliza la Definición 3.1 y la Definición 3.2, además de la ayuda del experto. Se fijamos C^{Nue} como *Empresa D* e *Invertir* como la solución del caso en memoria. Veamos ahora el proceso de asignación de riesgo:

- Primer atributo: *Número de años*
  R_1*(Número de años = 10, Solución = Invertir) = Medio*. Invertir en una compañía que se creó hace 10 años no presenta un nivel de riesgo extremo. Por eso, el experto asignó valor *Medio*, al riesgo de aplicar la solución *Invertir* en una compañía de 10 años de vida.

- Segundo atributo: *Sector*
  R_2*(Sector = Distribución, Solución = Invertir) = Medio*. El riesgo de aplicar la solución *Invertir* para una empresa del sector *Distribución* es *Medio*, no es un dato que aporte mucha información en este sentido.

- Tercer atributo: *Beneficio medio de los últimos 5 años*
  R_3*(Beneficio medio de los últimos 5 años = 1.5 %, Solución = Invertir) = Bajo*. El riesgo para la solución fijada es *Bajo*, porque invertir en una compañía cuyo beneficio medio en los últimos cinco años es positivo no conlleva riesgo asociado, ya que, el hecho de que genere beneficios indica que resulta fiable.

Tabla A.1: Tabla de Asignación de Riesgos: muestra el valor del riesgo en los atributos con valor crítico

	Años	**Sector**	**Beneficio**	**Cotiza**	**Empleados**	**Cash-flow**
Invertir	-	-	Bajo	-	-	Alto
No-Invertir	Bajo	Bajo	Bajo	Bajo	Bajo	Bajo

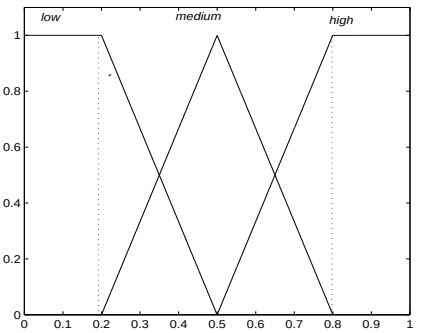
- Cuarto atributo: *Cotiza en bolsa*
  R_4*(Cotiza en bolsa = No, Solución = Invertir) = Medio*. Este dato no es determinante, hay empresas en las que pueda resultar interesante invertir pese a no haber salido a bolsa. Por lo tanto el experto asigna riesgo *Medio*.

- Quinto atributo: *Número de empleados*
  R_5*(Número de empleados = 2500, Solución = Invertir) = Medio*. Es una empresa lo suficientemente grande como para que este dato no produzca riesgo.

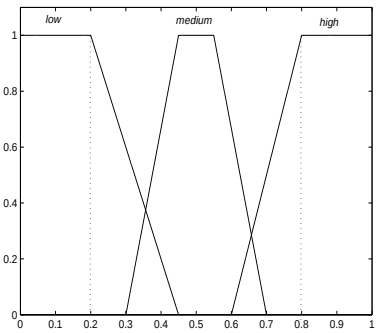
- Sexto atributo: *Cash-flow medio de los últimos 5 años*
  R_6*(Cash-flow medio de los últimos 5 años* $= -1.3$*, Solución = Invertir) = Alto*. El riesgo de aplicar la solución Invertir para este atributo cuyo valor es -1.3 es *Alto*, debido a que invertir en una compañía cuyo cash-flow en los últimos años ha sido negativo, es un indicativo que la compañía podría no ser rentable.

Se fija ahora *No-Invertir* como solución del caso en memoria y se repite el procedimiento. La Tabla A.1 muestra los resultados, aunque no se indican todos debido a que en la práctica estamos interesados solo en los casos críticos. Se entiende como caso crítico aquel atributo con valor de riesgo alto o bajo.

Una vez están asignados todos los riesgos, empieza la búsqueda del caso más adecuado. Para ello, el siguiente paso será asignar etiquetas lingüísticas tanto a los valores de la similitud local como a los pesos y así poder inferir en el sistema. En este ejemplo no es necesario asignar etiquetas lingüísticas al riesgo, puesto que ha sido asignado directamente con valores lingüísticos. Esto no supone ninguna restricción, puesto que si los valores son asignados en forma numérica, se procedería igual que con los pesos y similitudes. En la Figura A.1(a) y en la Figura A.1(b) pueden verse las funciones de pertenencia asociadas a la similitud y al peso respectivamente.



(a) Función de pertenencia para la similitud



(b) Función de pertenencia para los pesos

Figura A.1: Funciones de pertenencia

Veamos ahora como se disparan las reglas, para ello se elige un caso de la base de casos, por ejemplo, *Empresa A* y un atributo, por ejemplo el tercer atributo, *Beneficio medio de los últimos 5 años*, cuyo valor es 3.25, su peso $\omega_3 = 0.63$ y su similitud $sim(x_3^{Empresa\ A}, x_3^{Empresa\ D}) = 0.6759$. Las etiquetas lingüísticas para el peso, ω_3, y para la similitud, $sim(x_3^{Empresa\ A}, x_3^{Empresa\ D})$, pueden verse en las Figura A.1(b) y Figura A.1(a) respectivamente. El riesgo del atributo 3 para el caso *Empresa D* puede verse en la Tabla A.1, $R_3^{Inv} = bajo$. Con estos valores se disparan las siguientes reglas:

Si $R_3^{Inv} = bajo$, $\omega_3 = alto$ y $sim(x_3^{Empresa\ A}, x_3^{Empresa\ D}) = alto$, se dispara la Regla 11.

Si $R_3^{Inv} = bajo$, $\omega_3 = alto$ y $sim(x_3^{Empresa\ A}, x_3^{Empresa\ D}) = medio$, se dispara la Regla 12.

Si $R_3^{Inv} = bajo$, $\omega_3 = medio$ y $sim(x_3^{Empresa\ A}, x_3^{Empresa\ D}) = alto$, se dispara la Regla 14.

Si $R_3^{Inv} = bajo$, $\omega_3 = medio$ y $sim(x_3^{Empresa\ A}, x_3^{Empresa\ D}) = medio$, se dispara la Regla 15.

La fuerza con que se disparan las reglas se calcula con la Ecuación 3.6 para todas las reglas, salvo para la *Regla 10*, que por tratarse de una regla especial se usa la Ecuación 3.7. Por ejemplo, cuando se dispara *Regla 11* se calcula la fuerza del siguiente modo: la etiqueta lingüística bajo para el riesgo tiene un grado de pertenencia de 1, puesto que ha sido asignado directamente con un valor lingüístico. Para el peso su grado de pertenencia es $\mu_{alto}(0.63) = 0.15$, puede verse en la Figura A.1(b). Análogamente, el grado de pertenencia de la similitud es $\mu_{alto}(0.6759) = 0.58$. Utilizando la Ecuación 3.6, se obtiene:

$$g_{11} = \mu_{bajo}(R_3) \cdot \mu_{alto}(\omega_3) \cdot \mu_{alto}(sim(x_3^{Empresa\ A}, x_3^{Empresa\ D})) = 1 \cdot 0.15 \cdot 0.58 = 0.087$$

Se calculan las demás valores siguiendo el mismo procedimiento. La Tabla A.2 muestra todas las reglas que se han disparado y la fuerza con que lo han hecho, cada grupo de atributos se corresponde con un caso, el primero con el caso *Empresa A* el segundo con el caso *Empresa B* y el tercero con el caso *Empresa C*. La *Regla 10* se dispara cuando el atributo no tiene riesgo asociado. Los casos que no tienen riesgo asociado son aquellos a los que el experto no consideró situaciones de riesgo críticas y por tanto les asignó valor medio. En estos casos la fuerza de disparo de las reglas siempre es uno debido a la Ecuación 3.7 y a que el riesgo fue asignado directamente como con valor lingüístico.

Finalmente, se calcula la adecuación entre los casos *Empresa D* y *Empresa A*, *Empresa D* y *Empresa B* y también *Empresa D* y *Empresa C*. A continuación pueden verse los cálculos detallados:

Sim(Empresa A, Empresa D)$= \frac{1}{2.91}\{1 \cdot 0.0563 + 0 + \frac{1}{0.6105} \cdot [0.087 \cdot (0.4258 + 0.4 \cdot 0.4258) + 0.0615 \cdot (0.4258 + 0.3 \cdot 0.4258) + 0.27 \cdot (0.4258 + 0.2 \cdot 0.4258) + 0.192 \cdot (0.4258 + 0.2 \cdot 0.4258)] + 0 + 1 \cdot 0.0195 + 1 \cdot 0.1561\} = \frac{1}{2.91} \cdot (0.0563 + 0.526 + 0.0195 + 0.1561) = 0.26.$

Sim(Empresa B, Empresa D)$= \frac{1}{2.91}\{1 \cdot (0.4636 + 0.2 \cdot 0.4636) + 0 + frac10.621 \cdot [0.062 \cdot (0.2041 + 0.3 \cdot 0.2041) + 0.088 \cdot (0.2041 + 0.2 \cdot 0.2041) + 0.195 \cdot (0.2041 + 0.2 \cdot 0.2041) + 0.276 \cdot (0.2041 + 0.1 \cdot 0.2041)] + 1 \cdot (0.2 + 0.3 \cdot 0.2) + \frac{1}{0.68} \cdot [0.2 \cdot (0.31 + 0.2 \cdot 0.31) + 0.48 \cdot (0.31 + 0.3 \cdot 0.31)] + 1 \cdot (0.7695 + 0.4 \cdot 0.7695)\} = \frac{1}{2.91} \cdot (0.556 + 0.234 + 0.26 + 0.3939 + 1.0773) =$ **0.8663**.

Sim(Empresa C, Empresa D)$= \frac{1}{2.91} \cdot \{1 \cdot 0.3875 + 1 \cdot 0.42 + \frac{1}{0.62} \cdot [0.15 \cdot (0.5716 + 0.4 \cdot 0.5716) + 0.47 \cdot (0.5716 + 0.2 \cdot 0.5716)] + 1 \cdot 0.2 + 1 \cdot 0.3217 + 1 \cdot (0.0405 - 0.3 \cdot 0.0405)\} =$

Tabla A.2: Reglas disparadas y fuerza de disparo de cada una de ellas

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
A1										1									
A2										1									
A3											0.087	0.0615		0.27	0.192				
A4										1									
A5										1									
A6			1																
A1														1					
A2																0.8			0.12
A3												0.062	0.088		0.195	0.276			
A4																	1		
A5														0.2			0.48		
A6											1								
A1											1								
A2											1								
A3												0.15			0.47				
A4											1								
A5											1								
A6			1																

$\frac{1}{2.91} \cdot (0.3875 + 0.42 + 0.7135 + 0.2 + 0.3217 + 0.02835) = 0.7117.$

Los resultados, como cabía esperar, muestran que la *Empresa B* es el caso más similar, más adecuado, a la *Empresa D*.

Apéndice B

Estudio Gráfico de los Modelos de Asignación de Riesgo

En este apéndice se ilustra gráficamente a través de un ejemplo el comportamiento de cada uno de los métodos de asignación de riesgo que se estudiaron en el capítulo 4. Para continuar en la misma línea del ejemplo que se usó en el capítulo 3 al explicar la asignación de riesgo, se toma la base BUPA Liver Disorder, se fija el atributo Gamma-glutamil transpeptidasa (GGT) y la solución *No-Afectado*. La siguiente gráfica, Figura B.1 muestra todos los valores que toma ese atributo y su riesgo correspondiente dado por el experto, es decir, la verdadera función IRL dada por el experto.

Cada punto de la gráfica se interpreta de la siguiente forma, por ejemplo, el punto (200,1) significa que cuando el atributo GGT toma el valor 200 el experto le asignó un valor de riesgo de 1, si la solución que se considera era No-Afectado (el riesgo que la solución sea *No-Afectado* cuando el atributo toma el valor 200 es 1). Veamos ahora cómo es la aproximación que de la función IR, hace cada modelo. Se toman al azar 10 valores del atributo GGT y sus valores de riesgo correspondientes, ver Tabla B.1. Veamos ahora con estos datos de partida como es la aproximación proporcionada por cada método.

La aproximación mediante los *Modelos de Interpolación* puede verse en la Figura B.2, donde aparecen señalados todos los puntos de interpolación. Cada figura corresponde a un método propuesto. La Figura B.2(a) representa el método de Lagrange, como puede verse no es una buena aproximación debido a las oscilaciones que presenta al ser con 10 datos un polinomio de grado 9. Por la misma razón Hermite no

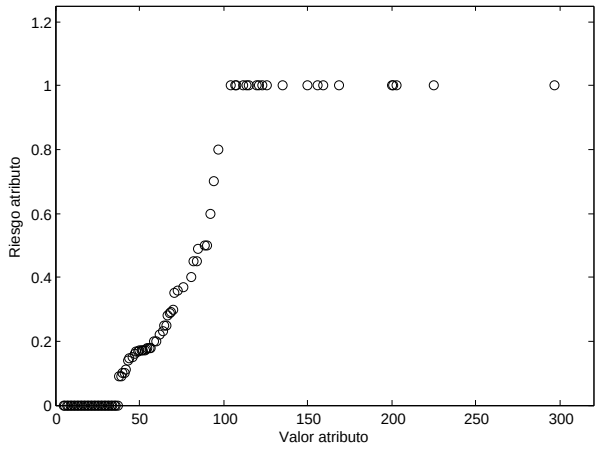


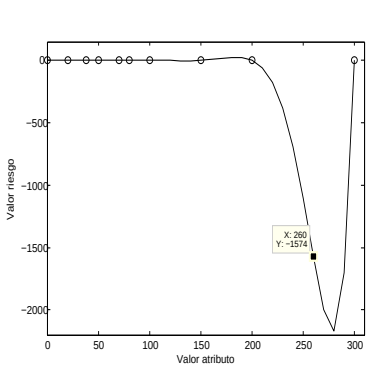
Figura B.1: Representación gráfica de la función Información de Riesgo dada por el experto para el atributo GGT cuando la solución considerada es No-Afectado

Tabla B.1: Datos para construir los ejemplos

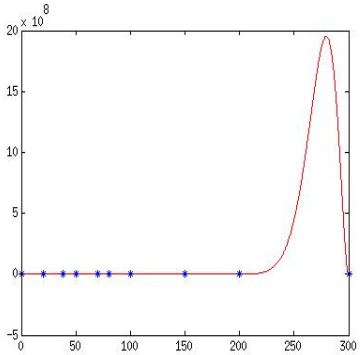
	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}
$GGT(x_i)$	0	20	38	50	70	80	100	150	200	300
$Riesgo(y_i)$	0	0	0.1	0.2	0.23	0.44	0.75	1	1	1

resulta adecuado (ver Figura B.2(b)). Finalmente puede verse como las funciones spline se ajustan bien al problema. La Figura B.2(c). muestra como resulta ser una buena aproximación a nuestro problema. Por ejemplo una predicción de riesgo para el valor 259 seria 0.9532 cuando sabemos que el valor que dio el experto fue 1. Los splines son numéricamente estables y además agradables a la vista, por lo que resulta un modelo muy interesante para aproximar como se ha visto en los resultados experimentales.

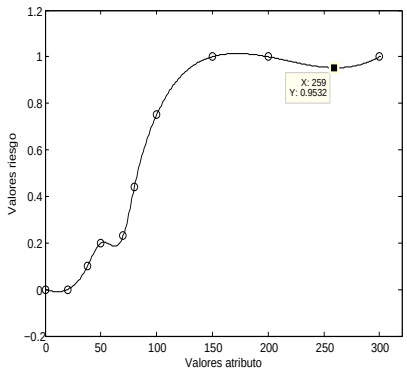
La Figura B.3 muestra gráficamente como se comportan los *Modelos de Aproximación*. En la Figura B.3(a) puede verse la recta regresión que según el criterio de mínimos cuadrados mejor se ajusta a los datos del ejemplo y la Figura B.3(b) la curva Bézier. Finalmente, la figura B.4 muestra cómo aproxima la información el método de Wang y Mendel considerando tres etiquetas difusas.



(a) Lagrage

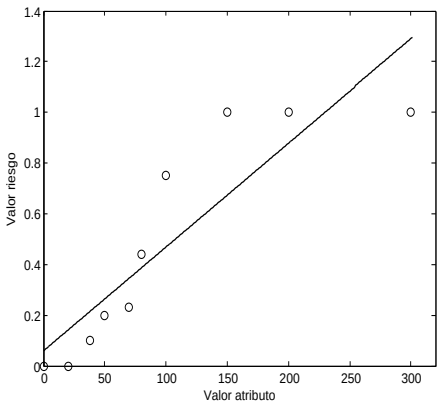


(b) Hermite

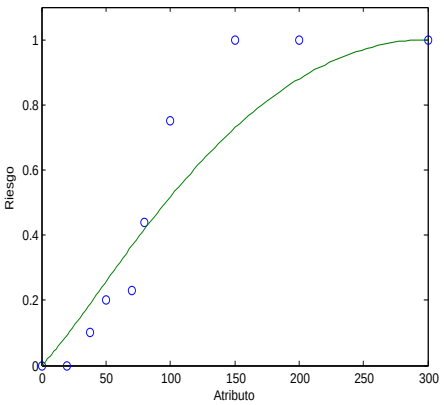


(c) Spline

Figura B.2: Modelos de Interpolación asociados al ejemplo de la Tabla B.1



(a) MCD



(b) Bézier

Figura B.3: Modelos de Aproximación asociados al ejemplo de la Tabla B.1

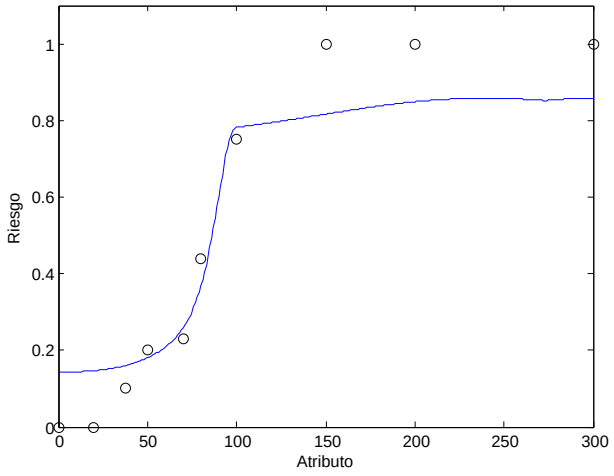


Figura B.4: Modelo de ajuste usando Wang y Mendel asociado al ejemplo de la Tabla B.1

Apéndice C

Cuestionario Utilizado en la Aplicación

La forma en que se recogieron los datos para construir el modelo de RBC encargado de estudiar la relación entre la flexibilidad y la estrategia de operaciones en una empresa, fue un cuestionario. Para diseñar el cuestionario, primero se envió a modo de prueba a 10 empresas con distinto tamaño y distintos campos de actividad. Esta prueba permitió corregir el enunciado de algunas cuestiones para así facilitar su comprensión. Cuando se tuvo la versión final del cuestionario se envió a 129 empresas cuya facturación fuese mayor de 150.000 € según datos de la Asociación Española de Empresas Consultoras en Ingeniería (Tecniberia/Asince). Se envió un e-mail al listado de empresas proporcionado por Tecniberia, solicitando colaboración. Con cada compañía se contactó al menos tres veces. Finalmente, solo 71 compañías (el 55 % del total) rellenaron el cuestionario. Una muestra del cuestionario final vacío puede verse en este apéndice.

CUESTIONARIO DE SERVICIOS PROFESIONALES DE CONSULTORÍA, INGENIERÍA Y SERVICIOS TECNOLÓGICOS

DEPARTAMENTO DE ORGANIZACIÓN DE EMPRESAS

UNIVERSIDAD COMPLUTENSE DE MADRID

Con la colaboración de: TECNIBERIA

Le agradecemos su colaboración en este proyecto que esperamos nos permita ampliar el conocimiento sobre la dirección de las operaciones en el sector, con el fin de que revierta en un futuro cercano, en las empresas del mismo. Si lo desea, existe una versión de este cuestionario a través de correo electrónico que puede solicitar simplemente enviando un mensaje con el encabezado CUESTIONARIO SERVICIOS a darias@ccee.ucm.es ♣

DATOS DE LA EMPRESA

Nombre de la empresa:_____
Dirección:_____
Teléfono:_____ Fax::_____ Dirección Internet: _____
Año de creación de la empresa:_____ Persona de contacto:_____

Facturación anual:
Menos de 50 millones ptas. Entre 50 y 100 millones ptas. De 100 a 500 millones ptas.
De 500 a 1.000 millones ptas. 1.000 o más millones ptas.

Número de empleados:
Menos de 10 Entre 10 y 99 100 o más

DATOS DEL ENCUESTADO

Cargo:_____
Unidad/Departamento al que pertenece:_____
Teléfono: _____ E-mail:_____

Formación (Marque con un círculo):
0.- Educación primaria
1.- Bachillerato/COU
2.- Diplomado/Ingeniero Técnico
3.- Licenciado/Ingeniero Superior
4.- Tercer ciclo (Master/Doctorado)

♣ Otras personas de contacto son:
- Jorge Mora Alberola (Tecniberia) Tfno.: (91) 431-37-60
- María José Álvarez Gil (Universidad Carlos III de Madrid) Tfno.: (91) 624-96-43. E-mail: catinaag@eco.uc3m.es
- Mariano Nieto Antolín (Universidad de León) . Tfno.: (987) 29-18-76. E-mail: ddemna@unileon.es

Bloque A. Estrategia de Operaciones

Por favor, marque con un círculo el número que mejor exprese su acuerdo o desacuerdo con las siguientes afirmaciones con relación a la función de operaciones de su empresa, según la siguiente escala:

1.- Absolutamente en desacuerdo. 2.- En desacuerdo aunque con reservas. 3.- Indiferente. 4.- De acuerdo aunque con reservas. 5.- Absolutamente de acuerdo

Bloque A.I. Distribución en planta

1) Las actividades relativas al proceso de prestación del servicio se realizan en un lugar preestablecido y fijo	1	2	3	4	5
2) Los recursos productivos están situados físicamente de manera secuencial	1	2	3	4	5
3) La distribución de los recursos para la prestación del servicio está realizada con fines de optimización del espacio y del propio proceso de prestación del servicio	1	2	3	4	5
4) Cada operario o trabajador está asignado exclusivamente a una tarea específica	1	2	3	4	5
5) Las tareas están secuenciadas de manera que no se comienza una hasta que no haya(n) terminado la(s) anterior(es)	1	2	3	4	5
6) En el diseño y distribución física del proceso priman los objetivos de eficiencia del sistema	1	2	3	4	5
7) Las actividades realizadas en el proceso de prestación del servicio se llevan a cabo en el lugar más conveniente para el cliente	1	2	3	4	5
8)Los recursos productivos pueden desplazarse al lugar o lugares donde se preste el servicio	1	2	3	4	5
9) La distribución física de los recursos se realiza con el fin de optimizar la prestación del servicio y la satisfacción del cliente	1	2	3	4	5
10) La asignación de los trabajadores a las tareas se realiza de manera rotatoria	1	2	3	4	5
11) En ocasiones los trabajadores realizan varias tareas distintas a la vez	1	2	3	4	5
12) En el diseño y distribución física del proceso priman los objetivos de satisfacción al cliente	1	2	3	4	5

Bloque A.II. Orientación PUSH/PULL del sistema

1) Se realizan grandes e importantes esfuerzos comerciales en la búsqueda y captación de nuevos clientes	1	2	3	4	5
2) Uno de los objetivos comerciales es que el cliente consuma el mayor número de servicios prestados por la empresa	1	2	3	4	5
3) Siempre se intenta aprovechar al máximo la capacidad productiva del sistema	1	2	3	4	5
4) Se intenta que el sistema de prestación del servicio funcione a su máximo nivel de eficiencia	1	2	3	4	5
5) Se realizan importantes esfuerzos en la realización de nuevas acciones que incrementen la satisfacción del cliente	1	2	3	4	5
6) El principal objetivo al prestar el servicio es la máxima satisfacción del cliente	1	2	3	4	5
7) Se prima la satisfacción del cliente por encima de la optimización de la capacidad	1	2	3	4	5

Bloque A.III. Nivel de estandarización

1) La mayor parte de las actividades del sistema de prestación del servicio están estandarizadas	1	2	3	4	5
2) El sistema de prestación del servicio está diseñado para que sólo exista una o muy pocas maneras de abordar cada tarea	1	2	3	4	5
3) Se intenta disminuir al máximo la posible variabilidad existente en las tareas	1	2	3	4	5
4) La mayoría de los procedimientos de trabajo están preestablecidos	1	2	3	4	5
5) El grado de autonomía e iniciativa concedido a los trabajadores es muy bajo	1	2	3	4	5
6) Todos aquellos sucesos no previstos en los procedimientos de trabajo han de ser comunicados a un superior para su resolución	1	2	3	4	5
7) Existe un libro de procedimientos y es conocido por los trabajadores	1	2	3	4	5
8) La mayoría de las actividades del sistema de prestación del servicio están orientadas a la adaptación del servicio al cliente	1	2	3	4	5
9) Existen numerosas formas diferentes de abordar cada tarea	1	2	3	4	5
10) La variabilidad de las tareas es alta	1	2	3	4	5
11) Los empleados gozan de un alto nivel de autonomía a la hora de establecer procedimientos de trabajo	1	2	3	4	5

Bloque A.IV. Abanico de servicios diferentes

1) La empresa ofrece un gran abanico de servicios diferentes	**1**	**2**	**3**	**4**	**5**
2) Los servicios ofrecidos se adaptan absolutamente a los deseos y necesidades de los clientes	**1**	**2**	**3**	**4**	**5**
3) Continuamente se diseñan y ofrecen nuevos servicios a los clientes	**1**	**2**	**3**	**4**	**5**
4) Los servicios ofrecidos a los clientes están estandarizados	**1**	**2**	**3**	**4**	**5**
5) La empresa presta uno, o muy pocos servicios diferentes, aunque muy especializados	**1**	**2**	**3**	**4**	**5**
6) La actividad de su empresa está dirigida a uno o pocos segmentos de clientes de tamaño pequeño o mediano	**1**	**2**	**3**	**4**	**5**

Bloque A.V. Uso de Tecnologías de la Información

1) El objetivo principal de la inversión en Tecnologías de la información es la reducción de costes	**1**	**2**	**3**	**4**	**5**
2) Se enfatiza la sustitución de mano de obra por nuevas tecnologías	**1**	**2**	**3**	**4**	**5**
3) Los clientes pueden enviar y/o recibir información sobre los servicios prestados a través de Tecnologías de la Información tales como Internet, Intercambio Electrónico de Datos (EDI), etc.	**1**	**2**	**3**	**4**	**5**
4) El objetivo principal de la inversión en Tecnologías de la información es la mejora del servicio al cliente	**1**	**2**	**3**	**4**	**5**
5) Las inversiones en Tecnologías de la Información tienen como objetivo el enriquecimiento de las tareas de los trabajadores	**1**	**2**	**3**	**4**	**5**
6) Las inversiones en Tecnologías de la Información tienen como objetivo la adaptación de los servicios al cliente	**1**	**2**	**3**	**4**	**5**
7) Se tiende a eliminar aquellos recursos que pueden ser sustituidos por la participación del cliente en el proceso	**1**	**2**	**3**	**4**	**5**
8) Las reorganizaciones del proceso se realizan siempre que mediante ellas disminuyan los costes	**1**	**2**	**3**	**4**	**5**

Bloque B. Flexibilidad del Sistema

Por favor, marque con un círculo el número que mejor exprese su acuerdo o desacuerdo con las siguientes afirmaciones con relación a su unidad de operaciones según la siguiente escala:
1.- Absolutamente en desacuerdo. 2.- En desacuerdo aunque con reservas. 3.- Indiferente. 4.- De acuerdo aunque con reservas. 5.- Absolutamente de acuerdo

1) La cantidad de tiempo necesaria para duplicar el output del sistema (servicios prestados) tiende a ser muy pequeña.	**1 2 3 4 5**
2) El coste de duplicar el output del sistema (servicios prestados) suele ser muy bajo.	**1 2 3 4 5**
3) La capacidad (por ejemplo, servicios prestados por unidad de tiempo) del sistema se puede incrementar con facilidad cuando sea necesario.	**1 2 3 4 5**
4) Las prestaciones (por ejemplo, la calidad) del sistema se pueden incrementar con facilidad cuando sea necesario.	**1 2 3 4 5**
5) La cantidad de tiempo necesario para la introducción de nuevos servicios es muy pequeña.	**1 2 3 4 5**
6) La cantidad de tiempo necesaria para incrementar la capacidad de servicio en una unidad es muy pequeña.	**1 2 3 4 5**
7) La habilidad de los sistemas informáticos para distribuir la información, procesarla y presentarla de la manera y en el momento adecuado a la persona que la solicite es muy alta.	**1 2 3 4 5**
8) El número de tareas diferentes que el sistema informático permite que se realicen en los ordenadores o terminales disponibles para el personal es muy alto.	**1 2 3 4 5**
9) El sistema informático permite intercambiar información de manera eficiente entre todos los ordenadores y terminales del sistema.	**1 2 3 4 5**
10) La disminución de la eficiencia del proceso de prestación del servicio debido a la avería de una máquina (ordenador, terminal, instrumento de medición, etc.) es muy baja.	**1 2 3 4 5**
11) La disminución de la eficiencia del proceso de prestación del servicio motivada por a la ausencia de un empleado experto es muy baja.	**1 2 3 4 5**
12) El coste que conlleva el error de un operario al iniciar el desarrollo de un proyecto es muy bajo.	**1 2 3 4 5**
13) Los equipos productivos pueden llevar a cabo operaciones muy distintas sin que ello conlleve unos costes muy altos de cambio de operación	**1 2 3 4 5**
14). Los equipos productivos pueden llevar a cabo operaciones muy distintas sin que ello conlleve unos tiempos muy altos de cambio de operación	**1 2 3 4 5**
15) Los trabajadores pueden llevar a cabo operaciones muy distintas sin que ello conlleve unos costes muy altos de cambio de operación	**1 2 3 4 5**
16). Los trabajadores pueden llevar a cabo operaciones muy distintas sin que ello conlleve unos tiempos muy altos de cambio de operación	**1 2 3 4 5**
17)El coste de dejar desatendido un pedido es muy bajo.	**1 2 3 4 5**
18) El coste de demora en un plazo de entrega ya pactado con el cliente es muy bajo.	**1 2 3 4 5**
19) El número de innovaciones introducidas en los servicios anualmente es muy alto.	**1 2 3 4 5**
20) El volumen de servicios que el sistema de operaciones es capaz de desarrollar sin añadir equipamiento es muy grande.	**1 2 3 4 5**
21) Un empleado puede llevar a cabo un número muy alto de tareas diferentes	**1 2 3 4 5**
22) En muy pocas ocasiones se encuentran servicios en curso en espera (stand-by) por falta de capacidad productiva.	**1 2 3 4 5**
23) El sistema de operaciones es capaz de trabajar sin supervisión durante, al menos, una jornada laboral.	**1 2 3 4 5**
24) El rango de volumen de servicio dentro del cual la empresa puede operar de manera rentable es muy alto.	**1 2 3 4 5**

Bloque C. Rendimiento del sistema

Por favor, indique para cada una de las medidas de rendimiento que se proponen a continuación, la importancia que tiene en su organización, así como el nivel de resultados logrados en cada una de ellas con relación a sus competidores. Resultaría de gran interés que añadiese al final de la tabla otras medidas que considere oportunas. (Marque con un círculo el número apropiado)

	Ninguna importancia	*Poca importancia*	*Cierta importancia*	*Gran importancia*	*Medida fundamental*	*Peor que la competencia*	*Algo peor que la competencia*	*Igual*	*Algo mejor que la competencia*	*Mejor que la competencia*
1) Volumen de servicios prestados	1	2	3	4	5	1	2	3	4	5
2) Número de quejas de clientes	1	2	3	4	5	1	2	3	4	5
3) Evolución del nivel de eficiencia de los trabajadores	1	2	3	4	5	1	2	3	4	5
4) Frecuencia de errores en la prestación del servicio	1	2	3	4	5	1	2	3	4	5
5) Tasa de beneficios en relación con los ingresos netos	1	2	3	4	5	1	2	3	4	5
6)Reducción de los tiempos de puesta a punto de los equipos	1	2	3	4	5	1	2	3	4	5
7) Grado de ajuste de los costes reales respecto a los costes estándar preestablecidos	1	2	3	4	5	1	2	3	4	5
8) Tasa de introducción de nuevos servicios	1	2	3	4	5	1	2	3	4	5
9) Rentabilidad de los fondos propios	1	2	3	4	5	1	2	3	4	5
10) Número de servicios iniciados pero no finalizados por petición del cliente	1	2	3	4	5	1	2	3	4	5
11) Variación en los precios de compra de materiales	1	2	3	4	5	1	2	3	4	5
12) Duración de tiempo transcurrido entre la petición del servicio por parte del cliente y la prestación final del mismo	1	2	3	4	5	1	2	3	4	5
13) Porcentaje de utilización de la capacidad productiva	1	2	3	4	5	1	2	3	4	5
14) Evaluación de la habilidad para variar, adaptar o personalizar las características del servicio	1	2	3	4	5	1	2	3	4	5
15) Porcentaje de tiempo productivo frente al ocioso	1	2	3	4	5	1	2	3	4	5
16) Nivel de satisfacción de los clientes con la empresa	1	2	3	4	5	1	2	3	4	5
17) Reducción de costes debida a incrementos en calidad	1	2	3	4	5	1	2	3	4	5
18) Nivel de cooperación de los trabajadores a nivel interdepartamental	1	2	3	4	5	1	2	3	4	5
	1	2	3	4	5	1	2	3	4	5
	1	2	3	4	5	1	2	3	4	5
	1	2	3	4	5	1	2	3	4	5
	1	2	3	4	5	1	2	3	4	5
	1	2	3	4	5	1	2	3	4	5
	1	2	3	4	5	1	2	3	4	5

Actividades que realiza la empresa:

Ingeniería	***Consultoría***
□ *Transporte y Comunicaciones*	□ *Procesos Industriales*
□ *Hidrología e Hidráulica*	□ *Sistemas de Información*
□ *Geología y Geotecnia*	□ *Recursos Humanos*
□ *Agronomía, Pesca y Ganadería*	□ *Financiera*
□ *Urbanismo y Arquitectura*	□ *Comercio Exterior*
□ *Abastecimiento y Saneamiento*	□ *I+D*
□ *Medio Ambiente*	□ *Análisis Contable*
□ *Energía*	□ *Marketing*
□ *Minería*	□ *Servicios tecnológicos*
□ *Plantas Industriales*	□ *Otras (Indicar):*
□ *Plantas Químicas*	□
□ *Cartografía, Topografía y Fotogrametría*	□
□ *Estudios Macroeconómicos y Sociales*	□

Este espacio queda reservado para cualquier sugerencia, comentario o aportación que desee realizar respecto al estudio. De nuevo le reiteramos nuestro agradecimiento por su colaboración.

Bibliografía

[1] A. Aamodt. Towards robust expert systems that learn from experience an architectural framework. In *Proceedings of the third European Workshop on Knowledge-Based Systems (EKAW)*, pages 311–326, 1989.

[2] A. Aamodt. *A Knowledge-intensive, Integrated Approach to Problem Solving and Sustained Learning*. PhD thesis, University of Trondheim, 1991.

[3] A. Aamodt and E. Plaza. Case-based reasoning: foundational issues, methodological variations and system approaches. *AI Communications*, 7(1):39–59, 1994.

[4] H. Ahn and K.-J. Kim. Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing*, 9(2):599–607, 2009.

[5] H. Ahn and K.-J. Kim. Global optimization of case-based reasoning for breast cytology diagnosis. *Expert Systems with Applications*, 36(1):724–734, 2009.

[6] H. Ahn, K.-J. Kim, and I. Han. A case-based reasoning system with the two-dimensional reduction technique for customer classification. *Expert Systems with Applications*, 32(4):1011–1019, 2007.

[7] M. Allais and O. Hagen. *Expected utilility Hypotheses and the Allais paradox*. Reidel, Dordrecht, 1979.

[8] K.D. Althoff, B. Faupel, S. Kockskämper, R. Traphöner, and W. Wernicke. Knowledge acquisition in the domain of cnc machining centers: the moltke approach. In *Proceedings of the third European Workshop on Knowledge-Based Systems (EKAW)*, pages 180–195, 1989.

[9] D. Arias-Aranda. Service operations strategy, flexibility and performance in engineering consulting firms. *International Journal of Operations and Production Management*, 23(11):1401–1424, 2003.

[10] D. Arias-Aranda, J.L. Castro, M. Navarro, J.M. Sánchez, and J.M. Zurita. A fuzzy expert system for business management. *Expert Systems with Applications*, 37(12):7570– 7580, 2010.

[11] D. Arias-Aranda, J.L. Castro, M. Navarro, and J.M. Zurita. A cbr system for knowing the relationship between flexibility and operations strategy. In *Lecture Notes in Computer Science*, pages 463–472, 2009.

[12] E. Armengol and E. Plaza. Relational case-based reasoning for carcinogenic activity prediction. *Artificial Intelligence Review*, 20(1):121–141, 2003.

[13] K.D. Ashley and E.L. Rissland. A case-based approach to modeling legal expertise. *IEEE Expert: Intelligent Systems and Their Applications*, 3(3):70–77, 1988.

[14] T. Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions*, 53:370–418, 1783.

[15] G.M. Becker, M.H. DeGroot, and J. Marschak. Measuring utility by a single-response sequential method. *Behav. Sci.*, 9:226–232, 1964.

[16] S. Begum, M.U. Ahmed, P. Funk, N. Xiong, and B. Von Schele. A case-based decision support system for individual stress diagnosis using fuzzy similarity matching. *Computational Intelligence*, 25(3):180–195, 2009.

[17] R. Bergmann, J. Kolodner, and E. Plaza. Representation in case-based reasoning. *The Knowledge Engineering Review*, 20(3):209–213, 2006.

[18] I. Bichindaritz and C. Marling. Case-based reasoning in the health sciences: What's next? *Artificial Intelligence in Medicine*, 36(2):127–135, 2006.

[19] A. Bonzano, P. Cunningham, and B. Smyth. Using introspective learning to improve retrieval in cbr: a case study in air trafic control. In *Proceedings of the second International Conference on Case-Based Reasoning (ICCBR)*, pages 291–302, 1997.

[20] D. Bridge. Defining and combining symmetric and asymmetric similarity measures. In *Lecture Notes in Computer Science*, pages 52–63, 1998.

[21] A.J. Broder. Strategies for efficient incremental nearest neighbor search. *Pattern Recognition*, 23(1–2):171–178, 1990.

[22] F.J. Cabrerizo, I.J. Pérez, and E. Herrera-Viedma. Managing the consensus in group decision making in an unbalanced fuzzy linguistic context with incomplete information. *Knowledge-Based Systems*, 23(2):169–181, 2010.

[23] J.L. Castro, M. Navarro, K. Benghazi, M.V. Hurtado, and J.M. Zurita. A cbr system to new product development: An application for hearing devices design. In *Proceedings of World Academic of Science, Engineering and Technology Conference*, pages 295–300, 2010.

[24] J.L. Castro, M. Navarro, J.M. Sánchez, and J.M. Zurita. Similarity local adjustment: Introducing attribute risk into the case. In *Proceedings of the European and Mediterranean Conference on Information Systems (EMCIS)*, 2006.

[25] J.L. Castro, M. Navarro, J.M. Sánchez, and J.M. Zurita. Global risk attribute in case-based reasoning. In *Proceedings of the 7th International Conference on Case-Based Reasoning (ICCBR)*, pages 21–30, 2007.

[26] J.L. Castro, M. Navarro, J.M. Sánchez, and J.M. Zurita. An automatic method to assign local risk. In *Proceedings of the International Multi Conference on Computer Science and Information Systems (IADIS)*, pages 151–157, 2008.

[27] J.L. Castro, M. Navarro, J.M. Sánchez, and J.M. Zurita. Loss and gain functions for cbr retrieval. *Information Sciences*, 179(11):1738–1750, 2009.

[28] J.L. Castro, M. Navarro, J.M. Sánchez, and J.M. Zurita. Introducing attribute risk for retrieval in case-based reasoning. *Knowledge-Based Systems*, 24(2):257–268, 2010.

[29] C.L. Chang, B.W. Cheng, and J.L. Su. Using case-based reasoning to establish a continuing care information system of discharge planning. *Expert Systems with Applications*, 26(4):601–613, 2004.

[30] P.-C. Chang, C.-Y. Fan, and W.-Y. Dzan. A cbr-based fuzzy decision tree approach for database classification. *Expert Systems with Applications*, 37(1):214–225, 2010.

[31] P.-C. Chang, C.-H. Lai, and K.R. Lai. A hybrid system by evolving case-based reasoning with genetic algorithm in wholesaler's returning book forecasting. *Decision Support Systems*, 42(3):1715–1729, 2006.

[32] P.-C. Chang, J.-J. Lin, and W.-Y. Dzan. Forecasting of manufacturing cost in mobile phone products by case-based reasoning and artificial neural network models. *Journal of Intelligent Manufacturing*, pages 1–15, 2010.

[33] P.-C. Chang, C.H. Liu, and R.K. Lai. A fuzzy case-based reasoning model for sales forecasting in print circuit board industries. *Expert Systems with Applications*, 34(3):2049–2058, 2008.

[34] S. Chaudhury, T. Singh, and P.S. Goswami. Distributed fuzzy case based reasoning. *Applied Soft Computing*, 4(4):323–343, 2004.

[35] F. Chen, B. Zhang, and L. Bai. Design and realization of case-indexing model based on ahp. *Journal of Software*, 5(8):851–857, 2010.

[36] S.-M. Chen, M.-S. Yeh, and P.-Y. Hsiao. A comparison of similarity measures of fuzzy values. *Fuzzy Sets and Systems*, 72(1):79–89, 1995.

[37] W.C. Chen, S.S. Tseng, L.P. Chang, T.P. Hong, and M.F. Jiang. A parallelized indexing method for large-scale case-based reasoning. *Expert Systems with Applications*, 23(2):95–102, 2002.

[38] M.-Y. Cheng, H.-C. Tsai, and Y.-H. Chiu. Fuzzy case-based reasoning for coping with construction disputes. *Expert Systems with Applications*, 36(2):4106–4113, 2009.

[39] C.-C. Chiu and C.-Y. Tsai. A weighted feature c-means clustering algorithm for case indexing and retrieval in cased-based reasoning. In *Proceedings of the 20th International Conference on Industrial, engineering, and other applications of applied intelligent systems*, pages 541–551, 2007.

[40] C.-Y. Chiu, C.-C. Lo, and Y.-X. Hsu. Integrating bayesian theory and fuzzy logics with case-based reasoning for car-diagnosing problems. In *Proceedings of the fourth International Conference on Fuzzy Systems and Knowledge Discovery*, pages 344–348, 2007.

[41] J.P. Chou. *Multivariate exponential families: an identity and the admissibility od standard estimates*. PhD thesis, University of California, 1984.

[42] S. Conte. *Análisis numérico elemental un enfoque algoritmico.* Mcgraw-Hill, 1972.

[43] J.M. Corchado, J. Bajo, J.F. De Paz, and S. Rodríguez. An execution time neural-cbr guidance assistant. *Neurocomputing*, 72(13–15):2743–2753, 2009.

[44] T. Cover and P. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21–27, 1967.

[45] J. Daengdej, D. Lukose, E. Tsui, P. Beinat, and L. Prophet. Combining case-based reasoning and statistical method for proposing solution in ricad. *Knowledge-Based Systems*, 10(3):153–159, 1997.

[46] B.V. Dasarathy. *Nearest neighbor (NN) norms: NN pattern classification techniques.* Los Alamitos, California, IEEE Computer Society Press, 1991.

[47] J. de Miguel, L. Plaza, and B. Díaz-Agudo. Colibricook: A cbr system for ontology-based recipe retrieval and adaptation. ECCBR Workshops. Computer Cooking Contest., 2008.

[48] M.H. DeGroot. *Optimal Statistical Decisions.* McGraw-Hill, New York, 1970.

[49] S.J. Delany, P. Cunningham, and L. Coyle. An assessment of case-based reasoning for spam filtering. *Artificial Intelligent Review*, 24(3–4):359–378, 2005.

[50] L. Dengfenga and C. Chuntianb. New similarity measures of intuitionistic fuzzy sets and application to pattern recognitions. *Pattern Recognition Letters*, 23(1–3):221–225, 2002.

[51] S.Z. Dogan, D. Arditi, and H.M. Günaydin. Using decision trees for determining attribute weights in a case-based model of early cost prediction. *Journal of Construction Engrg. and Mgmt.*, 134(2):146–152, 2008.

[52] D. Dubois, H. Prade, and C. Testemal. Weighted fuzzy pattern matching. *Fuzzy Sets and Systems*, 28(3):313–331, 1988.

[53] L. Dymova and P. Sevastjanov. An interpretation of intuitionistic fuzzy sets in terms of evidence theory: Decision making aspect. *Knowledge-Based Systems*, 23(8):772–782, 2010.

[54] A.W.F. Edwards. The history of likelihood. *International Statistical Review*, 42(1):9–15, 1974.

[55] M. Elter, T. Wittenberg, and R. Schulz-Wendtland. A case-based reasoning system using feature scaling for computer aided breast cancer. *Computer-Assisted Radiology and Surgery*, 2:340–342, 2007.

[56] J.R. Evans. An exploratory study of performance measurement systems and relationships with performance results. *Journal of Operations Management*, 22(3):219–232, 2004.

[57] A.O. Finley and R.E. McRoberts. Efficient k-nearest neighbor searches for multi-source forest attribute mapping. *Remote Sensing of Environment*, 112(5):2203–2211, 2008.

[58] P.C. Fishburn. Subjetive expected utility: a reviewof normative theories. *Theory and Decision*, 13(2):139–199, 1981.

[59] G. Florez-Puga, B. Díaz-Agudo, and P. González-Caldero. Similarity measures in hierarchical behaviours from a structural point of view. In *Proceedings of the 23rd International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 330–335, 2010.

[60] A. Frank and A. Asuncion. Uci machine learning repository, 2010. http://archive.ics.uci.edu/ml.

[61] L. Fu, D.H. Goh, and S. Foo. Query clustering using a hybrid query similarity measure. *WSEAS Transaction on Computers*, 3(3):700–705, 2004.

[62] P. Gançarski, A. Blansché, and A. Wania. Comparison between two coevolutionary feature weighting algorithms in clustering. *Pattern Recognition*, 41(3):983–994, 2008.

[63] W.S. Gosset. *Letters from W. S. Gosset to R. A. Fisher, 1915-1936, With Summaries by R. A. Fisher and a Foreword by L. McMullen.* Printed for private circulation, Z.Z., 1970.

[64] M. Gu, X. Tong, and A. Aamodt. Comparing similarity calculation methods in conversational cbr. In *Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI)*, pages 427–432, 2005.

[65] S. Ha. A personalized counseling system using case-based reasoning with neural symbolic feature weighting (cansy). *Applied Intelligence*, 29(3):279–288, 2008.

[66] R.W. Hamming. Error detecting and error correcting codes. *Bell System Technical Journal*, 29(2):147–160, 1950.

[67] K. Hanney and M.T. Keane. Learning adaptation rules from a case-base. In *Lecture Notes in Computer Science*, pages 179–192, 1996.

[68] F. Hartge, W.E T. Wetter, and Haefeli. A similarity measure for case based reasoning modeling with temporal abstraction based on cross-correlation. *Computer Methods and Programs in Biomedicine*, 81(1):41–48, 2006.

[69] T.R. Hinrichs. *Problem solving in open worlds: A Case Study in Desing.* Lawrence Erlbaum Associates, 1992.

[70] A. Hoffmann and A.S. Khan. A new approach for the incremental development of retrieval functions for cbr. *Applied Artificial Intelligence*, 20(6):507–542, 2006.

[71] C.-C. Hsu and C.-S. Ho. A new hybrid case-based architecture for medical diagnosis. In *Proceedings of the Joint Conference on Information Sciences*, pages 168–175, 2002.

[72] C.-C. Hsu and C.-S. Ho. A new hybrid case-based architecture for medical diagnosis. *Information Sciences*, 166(1–4):231–247, 2004.

[73] B.W. Huang, M.L. Shih, N.-H. Chiu, W.Y. Hu, and C. Chiu. Price information evaluation and prediction for broiler using adapted case-based reasoning approach. *Expert Systems with Applications*, 36(2):1014–1019, 2009.

[74] M.-J. Huang, H.-S. Huang, and M.-Y. Che. Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach. *Expert Systems with Applications*, 33(3):551–564, 2007.

[75] J.E. Hunt, D.E. Cooke, and H. Holstein. Case memory and retrieval based on the immune system. In *Proceedings of the first International Conference on Case-based Reasoning (ICCBR)*, pages 205–216, 1995.

[76] K.H. Im and S.C. Park. Case-based reasoning and neural network based expert system for personalization. *Expert Systems with Applications*, 32(1):77–85, 2007.

[77] P. Jaccard. Etude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[78] J. Jarmulak, S. Craw, and R. Rowe. Genetic algorithms to optimise cbr retrieval. In *Lecture Notes in Computer Science*, pages 136–147, 2000.

[79] Y.-K. Juan, S.-G. Shih, and Y.-H. Perng. Decision support for housing customization: A hybrid approach using case-based reasoning and genetic algorithm. *Expert Systems with Applications*, 31(1):83–93, 2006.

[80] J.M. Juarez, F. Guil, J. Palma, and R. Marin. Temporal similarity by measuring possibilistic uncertainty in cbr. *Fuzzy Sets and Systems*, 160(2):214–230, 2009.

[81] J.M. Juárez, J. Salort, J. Palma, and R. Marin. Case representation ontology for case retrieval systems in medical domains. In *Proceedings of the 25th International Multi-Conference: artificial intelligence and applications (IASTED)*, pages 168–173, 2007.

[82] M.T. Keane. Analogical asides on case-based reasoning. In *Lectures Notes in Computer Sciencies*, pages 21–32, 1994.

[83] M.T. Keane. *Analogical Problem Solving*. Ellis Horwood, Chichester, West Sussex, England, 1998.

[84] R.L. Keeney and H. Raiffa. *Decisions with multiple objectives*. Wiley, New York, 1976.

[85] K.-J. Kim. Toward global optimization of case-based reasoning systems for financial forecasting. *Applied Intelligence*, 21(3):239–249, 2004.

[86] K.-S. Kim and I. Han. The cluster-indexing method for case-based reasoning using self-organizing maps and learning vector quantization for bond rating cases. *Expert Systems with Applications*, 21(3):147–156, 2001.

[87] J. Kolodner. Maintaining organization in a dynamic long-term memory. *Cognitive Science: A Multidisciplinary Journal*, 7(4):243–280, 1983.

[88] J. Kolodner. Reconstructive memory: A computer model. *Cognitive Science: A Multidisciplinary Journal*, 7(4):281–328, 1983.

[89] J. Kolodner. *Case-Based Reasoning*. Morgan Kauffman, 1993.

[90] P. Kontkanen, J. Lathinen, P. Myllymäki, and H. Tirri. An unsupervised bayesian distance measure. In *Lecture Notes in Computer Science*, pages 148–160, 2000.

[91] P. Koton. *Using experience in learning and problem solving*. PhD thesis, Massachusetts Institute of Technology, Laboratory of Computer Science, 1989.

[92] S. Ku and Y.-H. Suh. An investigation of the k-tree search algorithm for efficient case representation and retrieval. *Expert Systems with Applications*, 11(4):571–581, 1996.

[93] K.A. Kumar, Y. Singh, and S. Sanyal. Hybrid approach using case-based reasoning and rule-based reasoning for domain independent clinical decision support in icu. *Expert Systems with Applications*, 36(1):65–71, 2009.

[94] R.J. Kuo, Y.P. Kuo, and K.-Y. Chen. Developing a diagnostic system through integration of fuzzy case-based reasoning and fuzzy ant colony system. *Expert Systems with Applications*, 28(4):783–797, 2005.

[95] J.Z.C. Lai, Y.-C. Liaw, and J. Liu. Fast k-nearest-neighbor search based on projection and triangular inequality. *Pattern Recognition*, 40(2):351–359, 2007.

[96] D.B. Leake. *Case-based reasoning experiences, lessons future directions*. Mit Press, Cambridge, Massachusetts, 1996.

[97] D.B. Leake and J. Kendall-Morwick. Four heads are better than one: Combining suggestions for case adaptation. In *Lectures Notes in Computer Sciencies*, pages 165–179, 2009.

[98] D.B. Leake, A. Kinley, and D. Wilson. Learning to integrate multiple knowledge sources for case-based reasoning. In *Proceedings of the fifteenth International Joint Conference on Artificial Intelligence*, pages 674–679, 1997.

[99] D.B. Leake and D. Wilson. Categorizing case-base maintenance: dimensions and directions. In *Proceedings of the fourth European Workshop on Case-Based Reasoning (EWCBR)*, pages 197–207, 1998.

[100] D.B. Leake and D. Wilson. Remembering why to remember: Performance-guided case-base maintenance. In *Lecture Notes in Computer Science*, pages 83–99, 2000.

[101] E.L. Lehmann. Student and small-sample theory. *Statistical Science*, 14(4):418–426, 1999.

[102] H. Li and J. Sun. Ranking-order case-based reasoning for financial distress prediction. *Knowledge-Based Systems*, 21(8):868–878, 2008.

[103] H. Li and J. Sun. Majority voting combination of multiple case-based reasoning for financial distress prediction. *Expert Systems with Applications*, 36(3):4363–4373, 2009.

[104] S.-T. Li and H.-F. Ho. Predicting financial activity with evolutionary fuzzy case-based reasoning. *Expert Systems with Applications*, 36(1):411–422, 2009.

[105] T. Warren Liao, Z. Zhang, and C.R. Mount. Similarity measures for retrieval in case-based reasoning systems. *Applied Artificial Intelligence*, 12(4):267–288, 1998.

[106] T.W. Liao. An investigation of a hybrid cbr method for failure mechanisms identification. *Engineering Applications of Artificial Intelligence*, 17(1):123–134, 2004.

[107] T.W. Liao and Z.M. Zhang. A review of similarity measures for fuzzy systems. In *Proceedings of the fifth IEEE International Conference on Fuzzy System*, pages 8–11, 1996.

[108] R.-H. Lin, Y.-T. Wang, C.-H. Wu, and C.-L. Chuang. Developing a business failure prediction model via rst, gra and cbr. *Expert Systems with Applications*, 36(2):1593–1600, 2009.

[109] C.-H. Liu, L.-S. Chen, and C.-C. Hsu. An association-based case reduction technique for case-based reasoning. *Information Sciences*, 178(17):3347–3355, 2008.

[110] C.R. Marling, G.J. Petot, and L.S. Sterling. Integrating cased-based and rule-based reasoning to meet multiple design constraints. *Computational Intelligence*, 15(3):308–332, 1999.

[111] V. Marques, J.T. Farinha, and A. Brit. Case-based reasoning and fuzzy logic in fault diagnosis. *WSEAS Transactions on Computers*, 8(8):1408–1417, 2008.

[112] F. Martinez and L. Garrido. Application of machine learning techniques for reducing retrieval time in large case-based reasoning systems. In *Proceedings of the 24th International Conference on Artificial Intelligence and Applications (IASTED)*, pages 287–292, 2006.

[113] D.L. Medin, R.L. Goldstone, and D. Gentner. Respects for similarity. *Psychological Review*, 100(2):254–278, 1993.

[114] S.J. Miah, D.V. Kerr, and J.G. Gammack. A methodology to allow rural extension professionals to build target-specific expert systems for australian rural business operators. *Expert Systems with Applications*, 36(1):735–744, 2009.

[115] C. Milare and A. Decarvalho. Using a neural network in a cbr system. In *Proceedings of the International Conference on Computacional Intelligence and Multimedia Applications*, pages 43–48, 1998.

[116] S. Miyamoto. Information retrieval based on fuzzy associations. *Fuzzy Sets and Systems*, 38(2):191–205, 1990.

[117] G.L. Murphy and D.L. Medin. The role of theories in conceptual coherence. *Psychological Review*, 92(2):289–316, 1995.

[118] J.F. Nash. The bargaining problem. *Econometrica*, page 18(2), 1950.

[119] N. Neagu and B. Faltings. Exploiting interchangeabilities for case adaptation. In *Lecture Notes in Computer Science*, pages 422–436, 2001.

[120] S. Negny and J.-M. Le Lann. Acceleration of the retrieval of past experiences in case based reasoning: application for preliminary design in chemical engineering. *Computer Aided Chemical Engineering*, 25:1009–1014, 2008.

[121] H. Núñez, M. Sánchez-Marre, U. Cortés, J. Comas, M. Martínez, I. Rodríguez-Roda, and M. Poch. A comparative study on the use of similarity measures in case-based reasoning to improve the classification of environmental system situations. *Environmental Modelling and Software*, 19(9):809–819, 2004.

[122] K.J. Oh and T.Y. Kim. Financial market monitoring by case-based reasoning. *Expert Systems with Applications*, 32(3):789–800, 2007.

[123] H.R. Osborne and D. Bridge. A case-based similarity framework. In *Proceedings of third European Workshop on Case-based Reasoning (EWCBR)*, pages 309–323, 1996.

[124] H.R. Osborne and D. Bridge. Similarity metrics: a formal unification of cardinal and non-cardinal similarity measures. In *Lecture Notes in Computer Science*, pages 235–244, 1997.

[125] S.K. Pal and S.C.K. Chiu. *Foundations of Soft Case-Based Reasoning*. Wiley, 2004.

[126] C.P. Pappis and N.I. Karacapilidis. A comparative assessment of measures of similarity of fuzzy values. *Fuzzy Sets and Systems*, 56(2):71–174, 1993.

[127] C.-S. Park and I. Han. A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications*, 23(3):255–264, 2002.

[128] Y.J. Park, B.C. Kim, and S.H. Chum. New knowledge extraction technique using probability for case-based reasoning: application to medical diagnosis. *Expert Systems*, 23(1):2–20, 2006.

[129] S. Passone, P.W.H. Chung, and V. Nasseh. Incorporating domain-specific knowledge into a genetic algorithm to implement case-based reasoning adaptation. *Knowledge-Based Systems*, 19(3):192–201, 2006.

[130] J.C.M. Peng. Simultaneous estimation of the parameters of independent poisson distribution. Technical report, Department of Statistics, Stanford University, 1975.

[131] P. Perner. Case-base maintenance by conceptual clustering of graphs. *Engineering Applications of Artificial Intelligence*, 19(4):381–393, 2006.

[132] B.W. Porter and E.R. Bareiss. Protos: An experiment in knowledge acquisition for heuristic classification tasks. Technical report, University of Texas at Austin, 1986.

[133] G. Quellec, M. Lamard, L. Bekri, G. Cazuguel, B. Cochener, and C. Roux. Multimodal medical case retrieval using decision trees. *ITBM-RBM*, pages 35–43, 2008.

[134] J.R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[135] B. Raphael, B. Domer, S. Saitta, and I.F.C. Smith. Incremental development of cbr strategies for computing project cost probabilities. *Advanced Engineering Informatics*, 21(3):311–321, 2008.

[136] E.B Reategui, J.A Campbell, and B.F Leao. Combining a neural network with case-based reasoning in a diagnostic system. *Artificial Intelligence in Medicine*, 9(1):5–27, 1997.

[137] J. Renauda, E. Levratb, and C. Fonteixc. Weights determination of owa operators by parametric identification. *Mathematics and Computers in Simulation*, 77(5–6):499–511, 2008.

[138] M.M. Richter and S. Weiss. Uncertainty and case-based reasoning in patdex, 1991.

[139] E.L. Rissland and D.B. Skalak. Cabaret: rule interpretation in a hybrid architecture. *International Journal of Man-Machine Studies*, 34(6):839–887, 1991.

[140] H. Rubin. A weak system of axioms for rational"behavior and the nonseparability of utility from prior. *Statistics and Decisions*, 5:47–58, 1987.

[141] Santiago S. Ontañón and E. Plaza. Collaborative case retention strategies for cbr agents. In *Lecture Notes in Computer Science*, pages 1063–1063, 2003.

[142] A.W. Sadek, B.L. Smith, and M.J. Demetsky. A prototype case-based reasoning system for real-time freeway traffic routing. *Transportation Research Part C: Emerging Technologies*, 9(5):353–380, 2001.

[143] M.H. Safizadeh and D. Mallick L.P. Ritzman. Revisiting alternative theoretical paradigms in manufacturing strategy. *Production and Operations Management*, 9(2):111–127, 2000.

[144] R. Saraçolu, K. Tütüncü, and N. Allahverdia. A fuzzy clustering approach for finding similar documents using a novel similarity measure. *Expert Systems with Applications*, 33(3):600–605, 2007.

[145] J.W. Schaaf. Fish and shrink: a next step towards efficient case retrieval in large scaled case bases. In *Proceedings of the third European Workshop: Advances in Case-Based Reasoning*, pages 362–377, 1996.

[146] R.C. Schank. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People.* Cambridge University Press New York, NY, USA, 1983.

[147] R.C. Schank and R.P. Abelson. *Scripts, plans, goals and understanding: an inquiry into human knowledge structures.* Lawrence Erlbaum Associates, 1977.

[148] A. Schwering. Hybrid model for semantic similarity measurement. In *Lecture Notes in Computer Science*, pages 1449–1465, 2005.

[149] S. Sharma and D. Sleeman. Refiner: A case-based differential diagnosis aide for knowledge acquisition and knowledge refinement. In *Proceedings of European Working Session on Learning (EWSL)*, pages 201–210, 1988.

[150] K.S. Shin and I. Han. Case-based reasoning supported by genetic algorithms for corporate bond rating. *Expert Systems with Applications*, 12(2):85–95, 1999.

[151] S.C.K. Shiu, C.H. Sun, X.Z. Wang, and D.S. Yeung. Maintaining case-based reasoning systems using fuzzy decision tree. In *Proceedings of the 5th European workshop on case-based reasoning (EWCBR)*, pages 285–296, 2000.

[152] W. Shiue, S.-T. Li, and K.-J. Chen. A frame knowledge system for managing financial decision knowledge. *Expert Systems with Applications*, 35(3):1068–1079, 2008.

[153] M. Singh, M.K. Mandal, and A. Basu. Gaussian and laplacian of gaussian weighting functions for robust feature based tracking. *Pattern Recognition Letters*, 26(13):1995–2005, 2005.

[154] T.Y. Slonim and M. Schneider. Design issues in fuzzy case-based reasoning. *Fuzzy Sets and Systems*, 117(2):251–267, 2001.

[155] B. Smyth. Case-base maintenance. In *Lecture Notes in Computer Science*, pages 507–516, 1998.

[156] B. Smyth and M.T. Keane. Adaptation-guided retrieval: questioning the similarity assumption in reasoning. *Artificial Intelligence*, 102(2):249–293, 1998.

[157] B.-L. Su, M.-S. Wang, and Y.-M. Huang. Fuzzy logic weighted multi-criteria of dynamic route lifetime for reliable multicast routing in ad hoc networks. *Expert Systems with Applications*, 35(1–2):476–484, 2008.

[158] M.S Suh, W.C. Jhee, Y.K. Ko, and A. Lee. A case-based expert system approach for quality design. *Expert Systems with Applications*, 15(2):181–190, 1998.

[159] J. Sun and X.-F. Hui. Financial distress prediction based on similarity weighted voting cbr. In *Lecture Notes in Computer Science*, pages 947–958, 2006.

[160] Z. Sun, G. Finnie, and K. Weber. Case base building with similarity relations. *Information Sciences*, 165(1–2):21–43, 2004.

[161] E.P. Sycara. *Resolving adversarial conflicts: an approach integration case-based and analytic methods*. PhD thesis, Georgia Institute of Technology, 1987.

[162] E. Szmidt and J. Kacprzyk. A new concept of a similarity measure for intuitionistic fuzzy sets and its use in group decision making. In *Lecture Notes in Computer Science*, pages 272–282, 2005.

[163] T. Takagi and M. Sugeno. Fuzzy identification of systems and its application to modelling and control. *IEEE Transaction on Systems, Man and Cybernetics*, 15(1):116–132, 1985.

[164] C. Tsatsoulis, Q. Cheng, and H.Y. Wei. Integrating case-based reasoning and decision theory. *IEEE Intelligent Systems and Their Applications*, 12(4):46–55, 1997.

[165] K.-W. Tsui. Multiparameter estimation of discrete exponential distributions. *The Canadian Journal of Statistics*, 7(2):193–200, 1979.

[166] Y.-H. Tung, S.-S. Tseng, J.-F. Weng, T.-P. Lee, A.Y.H. Liao, and W.-N. Tsai. A rule-based cbr approach for expert finding and problem diagnosis. *Expert Systems with Applications*, 37(3):2427–2438, 2010.

[167] A. Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 1977.

[168] P. Viswanath, M. Narasimha, and S. Bhatnagar. Fusion of multiple approximate nearest neighbor classifiers for fast and efficient classification. *Information Fusion*, 5(4):239–250, 2004.

[169] A. Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

[170] D. Wang, Y. Xiang, G. Zou, and B. Zhang. Research on ontology-based case indexing in cbr. In *Proceedings of International Conference on Artificial Intelligence and Computational Intelligence*, pages 238–241, 2009.

[171] D.-G. Wang, Y.-P. Meng, and H.-X. Li. A fuzzy similarity inference method for fuzzy reasoning. *Computers & Mathematics with Applications*, 56(10):2445–2454, 2008.

[172] X. Wang, B. De Baets, and E. Kerre. A comparative study of similarity measures. *Fuzzy Sets and Systems*, 73(2):259–268, 1995.

[173] Y. Wang. Approximating nearest neighbor among triangles in convex position. *Information Processing Letters*, 108(6):379–385, 2008.

[174] I. Watson and F. Marir. Case-based reasoning : A review. *Knowledge Engineering Review*, 9(4):327–354, 1994.

[175] I. Watson and S. Perera. A hierarchical case representation using context guided retrieval. *Knowledge-Based Systems*, 11(5–6):285–292, 1998.

[176] I.D. Watson and S. Abdullah. Developing case-based reasoning systems: A case study in diagnosing building defects. In *Proceedings of IEEE Colloquium on Case-Based Reasoning: Prospects for Applications*, 1994.

[177] S. Weerahandi and J.V. Zidek. Elements of multi-bayesian decision theory. *The Annals of Statistics*, 11(4):1032–1046, 1983.

[178] S. Wess, K.-D. Althoff, and G. Derwand. Using k-d trees to improve the retrieval step in case-based reasoning. In *Lecture Notes in Computer Science*, pages 167–181, 1994.

[179] D. Wettschereck and D.W. Aha. Weithing features. In *Proceedings of the first International Conference on Case-based reasoning*, 1995.

[180] D. Wettschereck, D.W. Aha, and T. Mohri. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11:273–314, 1997.

[181] D.R. Wilson and T.R. Martinez. Improved heterogeneous distance functions. *Journal of Artificial Intelligence Research*, 6:1–34, 1997.

[182] I.H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco, 2005.

[183] F.S.Y. Wong, K.B. Chuah, and P.K. Venuvinod. Automated inspection process planning: Algorithmic inspection feature recognition, and inspection case representation for cbr. *Robotics and Computer-Integrated Manufacturing*, 22(1):56–68, 2006.

[184] D. Wu and J.M. Mendel. A vector similarity measure for linguistic approximation: Interval type-2 and type-1 fuzzy sets. *Information Sciences*, 178(2):381–402, 2008.

[185] M.-C. Wu, Y.-L. Lo, and S.-H. Hsu. A fuzzy cbr technique for generating product ideas. *Expert Systems with Applications*, 34(1):530–540, 2008.

[186] H. Xiaodong, W. Jianwu, S. Fuqian, and C. Haiyan. Apply fuzzy case-based reasoning to knowledge acquisition of product style. In *Proceeding of the IEEE 10th International Conference on Computer-Aided Industrial Design and Conceptual Design: E-Business, Creative Design, Manufacturing*, pages 383–386, 2009.

[187] Moreno Ribas y colaboradores. *Aprendizaje Automático*. Ediciones UPC, Universitat Politécnica de Cataluña, 1994.

[188] W. Yan, Q. Gao, Z. Liu, S. Zhang, and Y. Hu. A novel discovery technique on attribute weight of engine cbr design system. *Advanced Materials Research*, 97(101):3714–3717, 2010.

[189] C. Yang, B. Farley, and B. Orchard. Automated case creation and management for diagnostic cbr systems. *Applied Intelligence*, 28(1):17–28, 2008.

[190] S. Yang and D. Robertson. A case-based reasoning system for regulatory information. In *Proceedings of IEEE Colloquium on Case-Based Reasoning: Prospects for Applications*, 1994.

[191] Z. Yang, F. Deng, W. Liu, and Y. Fang. A cbr method for cfw prevention and treatment. *Expert Systems with Applications*, 36(3):5469–5474, 2009.

[192] Z. Yu, X. Zhou, L. Zhou, and K. Du. A hybrid similarity measure of contents for tv personalization. *Multimedia Systems*, 16(4):231–241, 2010.

[193] Zhongliang Yue. An extended topsis for determining weights of decision makers with interval numbers. *Knowledge-Based Systems*, 24(1):146–153, 2011.

[194] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[195] H.-Y. Zhang and W.-X. Zhang. Hybrid monotonic inclusion measure and its use in measuring similarity and distance between fuzzy sets. *Fuzzy Sets and Systemsl*, 160(1):107–118, 2009.

[196] F. Zhu, J. Guan, Y. Wang, and B. Liao J. Zhou. An automatic negotiation method based on cbr and agent reasoning. In *The Fifth International Conference on Computer and Information Technology*, 2005.

[197] Z.Y. Zhuang, L. Churilov, F. Burstein, and K.Sikaris. Combining data mining and case-based reasoning for intelligent decision support for pathology ordering by general practitioners. *European Journal of Operational Research*, 195(3):662–675, 2009.

[198] J.V. Zidek. Multi-bayesian estimation theory. *Statistal and Decisions*, 4:1–18, 1986.

[199] R. Zwick, E. Carlstein, and D.V. Budescu. Measures of similarity among fuzzy sets : A comparative analysis. *International Journal of Approximate Reasoning*, 1(2):221–242, 1987.