UNIVERSIDAD DE GRANADA

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

Departamento de Ciencias de la Computación e Inteligencia Artificial

MODELADO FORMAL PARA REPRESENTACIÓN Y EVALUACIÓN DE REGLAS DE ASOCIACIÓN

TESIS DOCTORAL

María Dolores Ruiz Jiménez

Granada, Enero de 2010

Editor: Editorial de la Universidad de Granada
Autor: María Dolores Ruiz Jiménez
D.L.: GR 2323-2010
ISBN: 978-84-693-1336-7

MODELADO FORMAL PARA REPRESENTACIÓN Y EVALUACIÓN DE REGLAS DE ASOCIACIÓN

MARÍA DOLORES RUIZ JIMÉNEZ

MODELADO FORMAL PARA REPRESENTACIÓN Y EVALUACIÓN DE REGLAS DE ASOCIACIÓN

MEMORIA QUE PRESENTA

MARÍA DOLORES RUIZ JIMÉNEZ

PARA OPTAR AL GRADO DE
DOCTOR EN INFORMÁTICA

Enero de 2010

DIRECTORES

MIGUEL DELGADO CALVO-FLORES
DANIEL SÁNCHEZ FERNÁNDEZ

DPTO. DE CIENCIAS DE LA COMPUTACIÓN E I.A.

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

UNIVERSIDAD DE GRANADA

La memoria titulada **Modelado Formal para Representación y Evaluación de Reglas de Asociación**, que presenta Dña. María Dolores Ruiz 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 D. Miguel Delgado Calvo-Flores y D. Daniel Sánchez Fernández.

Granada, Enero de 2010

El doctorando | Los directores

M.D. Ruiz Jiménez | M. Delgado Calvo-Flores | D. Sánchez Fernández

A mis padres, mi apoyo siempre.

A Paco, por estar ahí siempre.

"Ignorance is the curse of God,
Knowledge the wing wherewith we fly to heaven"
-*William Shakespeare*

Agradecimientos

El desarrollo de esta tesis no hubiera tenido lugar sin el apoyo de muchas personas a las que quiero agradecer su ayuda y dedicación.

En primer lugar quería dar mi gratitud a mis directores de tesis: Miguel Delgado y Daniel Sánchez. A Miguel por ser el guía durante todos estos años y por sus innumerables consejos y correcciones que han guiado mi investigación. A Daniel por todas las charlas que me han servido para reflexionar sobre las tareas que iba desarrollando y no sabía cómo darles forma. También por ayudarme a dar la importancia necesaria a todas las aportaciones que he ido dando en este mundo de la investigación.

También quiero agradecer a los miembros del departamento de Ciencias de la Computación e Inteligencia Artificial por darme su calurosa acogida como un miembro más de su familia, sin importar mi procedencia "matemática". En particular, quiero dar las gracias a los integrantes y los que alguna vez fueron integrantes de la Sala o Laboratorio 16, que me han soportado durante todo este tiempo y me han ayudado en mis dudas continuas en programación y otras cosas. Entre ellos quiero destacar algunos que han tenido alguna aportación especial en mi trabajo. A Pedro Magaña, Fernando Bobillo y Juan Gómez por su amistad, su inestimable ayuda en papeleos varios y en mi ausencia durante la estancia. A otros compañeros que me han ofrecido su amistad y siempre su ayuda para pequeñas dudas, por nombrar algunos: Carlos Cano, Javier López, Fernando García, Carlos Dancausa, Julián Garrido, Sergio Jaime y Eduardo Eisman, mi compañía en estos últimos días de escritura de la memoria. A Nacho García, Jesús Campaña y de nuevo a Pedro por su valiosa ayuda en todo el proceso de programación en java que he desarrollado. A otras personas que también me han brindado su amistad y me han dado buenas conversaciones durante estos años y han sido compañía en algunos

congresos: Aída Jiménez y María Ros. A Nacho Blanco por sus conversaciones y por toda su ayuda en temas docentes.

Desde el punto de vista económico, esta tesis no hubiera sido posible sin la remuneración ofrecida por el Ministerio de Educación y Ciencia a través de la beca de Formación de Profesorado Universitario (FPU) y las ayudas recibidas para las estancias en el extranjero desde el mismo Ministerio y desde el grupo de investigación TIC-111. También quiero agradecer al profesor Eyke Hüllermeier por su recepción en la Universidad de Marburg y su colaboración en nuestro reciente trabajo.

A mis amigos, muchos de ellos han aguantado más de una conversación sobre becas, plazas, docencia, artículos y demás, en especial a Elvira, Celia, Mari Carmen, Pilar y José Luis.

No me quiero olvidar de las personas más importantes que han estado presentes en todo momento. A mi madre Juani, por apoyarme y convencerme a estudiar matemáticas y a probar suerte en el mundo de la investigación. A mi padre Miguel, que aunque no está presente, sé que estaría contento de verme superar esta importante fase de mi vida. A Paco por estar a mi lado estos ocho años y medio y ofrecerme su apoyo en todas las decisiones que he ido tomando a lo largo de estos años. A mis hermanos, Miguel y Ana por aguantarme en casa los últimos tres años. Y a mi "chico", Francis, por ser la alegría de la familia.

Índice general

Índice de figuras

Índice de tablas

Introducción

En la actualidad el uso de los nuevos dispositivos para el almacenamiento de información va en creciente aumento. De forma paralela se han desarrollado diversas herramientas para obtener conocimiento de las distintas fuentes de información que posee el usuario. La información puede provenir de procedencias muy diferentes y puede estar estructurada de varias formas. La gran diversidad de formato de almacenaje de la información así como la gran variedad de su contenido, nos da mucha riqueza en cuanto al tipo de conocimiento que puede ser extraído a partir de una fuente de información.

A día de hoy, las bases de datos proporcionan una herramienta muy útil para estandarizar su almacenaje y partir de ellas se pueden desarrollar distintos mecanismos para encontrar información que sea útil, novedosa y de interés para el usuario. Estas tres características son muy importantes para poder discernir entre conocimiento y poder desechar lo que no aporte nada al usuario.

El proceso para la *Extracción del Conocimiento* se encarga de desarrollar todas las herramientas y procesos desde la selección, procesamiento y limpieza de los datos, hasta los mecanismos de *Minería de Datos* para obtener el conocimiento pasando por su interpretación y evaluación. Más concretamente una de las herramientas más utilizada en la *Minería de Datos* es la extracción de reglas de asociación debido a su sencilla interpretabilidad, aunque a veces hay que saber darle el significado justo y no otorgarles la causalidad que no poseen pudiendo confundir al usuario.

Este trabajo se centra en el desarrollo de un modelo formal para una mejor comprensión de los mecanismos de extracción y evaluación de las reglas de asociación, así como su desarrollo para unificar en un marco las distintas

propuestas que utilizan como herramienta las reglas de asociación o variantes de ellas. Además se han desarrollado transversalmente nuevas técnicas para obtener nuevos tipos de conocimiento desde distintas fuentes de información.

Antecedentes

El punto de partida de este trabajo es de un modelo llamado GUHA (General Unary Hypotheses Automaton) [Hájek et al., 1966] con origen en los años 60 para formular un método que permitiera a un ordenador generar y evaluar hipótesis que fueran interesantes desde el punto de vista de un problema concreto. La noción de hipótesis de GUHA es casi idéntica a la noción de regla de asociación introducida por Agrawal et al. en los años 90 [Agrawal et al., 1993]. Esta semejanza permite utilizar algunos de los fundamentos de GUHA para representar el funcionamiento y los procesos de evaluación de las reglas de asociación y ofrecer un nuevo punto de vista para su estudio.

Las reglas de asociación miden la posible relación que puede existir entre un conjunto de objetos en una base de datos midiendo su frecuencia de aparición en las transacciones. Usando dichas frecuencias se han propuesto distintas opciones para medir la validez, calidad y el interés del usuario en el conjunto de reglas extraídas en una base datos. Pocos han sido los esfuerzos por desarrollar un conjunto de axiomas o propiedades que deberían cumplir dichas medidas, y casi ninguna propuesta podemos encontrar para modelizar formalmente todo lo relacionado con las reglas de asociación. Esto se hace necesario para poder expresar numerosas formas de asociación que han surgido en estos años de investigación así como poder desarrollar nuevas con los formalismos necesarios. Lo que sí podemos encontrar son diversas teorías y modelos que dan lugar al desarrollo de nuevos algoritmos y nuevos tipos de reglas de asociación pero que no unifican todo el conocimiento ya desarrollado y existente sobre ellas.

Dos de los conceptos que serán básicos en el desarrollo del modelo serán la llamada tabla-4ft y el cuantificador. La tabla-4ft resume en una tabla o matriz de dimensión 2x2 las cuatro frecuencias en juego en una regla de asociación general del tipo $A \rightarrow B$, y el cuantificador va asociado a la medida de validez o cumplimiento de la regla de asociación. Según las propiedades y la forma de utilizar el cuantificador obtendremos distintos tipos de asociaciones con diferentes propiedades que pueden ser de interés para el usuario.

Además del desarrollo del modelo, en este trabajo presentaremos algunas

propuestas transversales para obtener nuevos tipos de conocimiento. Uno de ellos es un procedimiento para obtener varios tipos de asociaciones en una base de datos peculiar: en bolsas. Las bolsas, como su nombre indica, hacen referencia a un tipo especial de transacciones donde la principal diferencia radica en que los objetos tienen asociadas cantidades. Dichas cantidades hacen mención no sólo a la aparición del ítem en la transacción, si no, al número de veces que aparece. Esta información extra que ofrecen las bolsas puede aprovecharse de diversas formas para obtener asociaciones más ricas en conocimiento, por ejemplo asociaciones que involucran tendencias sobre el aumento o la disminución de las cantidades asociadas a los items de la regla.

Otra propuesta es el uso de la negación de los itemsets para obtener reglas de asociación con un conocimiento específico de utilidad para el usuario. Este tipo de reglas pueden ser infrecuentes pero su obtención se apoya en la búsqueda de pares de reglas, una de ellas frecuente y confidente, y la segunda, que involucra la negación, solamente confidente. Según la posición del ítem negado podemos obtener conocimiento que podemos asociar como del tipo excepcional o anómalo respecto de la regla frecuente y confidente. Los primeros trabajos sobre este tipo de reglas se propusieron a finales de los 90 y recientemente por diversos autores [Suzuki, 1996], [Hussain et al., 2000], [Berzal et al., 2004] y servirán como base a las propuestas que aquí presentamos.

Objetivos

La finalidad principal de este trabajo es el desarrollo de un modelo formal para la representación y evaluación de las reglas de asociación como herramienta en un proceso de minería de datos. Para ello hemos marcado unos objetivos y también se han desarrollado otros que son afines al principal propósito de un proceso de minería de datos: descubrimiento de conocimiento útil, novedoso y comprensible para el usuario.

A continuación describiremos los principales objetivos cubiertos con el desarrollo de este trabajo:

- Desarrollar un modelo formal para la representación y evaluación de las reglas de asociación.
  - Conocer los modelos existentes.
  - Reconocer un modelo y desarrollarlo para distintos tipos de reglas de asociación: crisp, difusas, de excepción y anómalas.

  - Estudiar las propiedades del modelo y establecer un modelo unificado.
  - Analizar un vínculo de unión entre los elementos del modelo y el tipo de asociación.
  - Obtener un conjunto de propiedades, principios o axiomas que debe tener una buena medida de interés en términos del modelo y estudiar condiciones para que las medidas para la evaluación de las reglas las satisfagan.

- Proporcionar nuevos procedimientos para obtener conocimiento de bases de datos formadas por bolsas.
  - Examinar las propuestas existentes en dicho ámbito.
  - Elaborar nuevos métodos de extracción que involucren y tengan en cuenta tanto si el objeto se encuentra o no en la base de datos, como la cantidad que tiene asociada.
  - Analizar y estudiar las ventajas e inconvenientes de dichos métodos.
  - Compararlos con los ya existentes.

- Analizar la semántica y encontrar nuevos métodos para la obtención de reglas de asociación que involucran la negación de items.
  - Realizar un amplio y riguroso estudio de los tipos de reglas de asociación propuestos hasta el momento que involucran la negación de items.
  - Ver cuáles recogen un mejor contenido semántico que pueda ser de interés para el usuario.
  - Modificar y/o proponer nuevos métodos y estrategias de extracción de los tipos de reglas que mejoren a los existentes.
  - Utilizar el modelo para su estudio y formalización.
  - Obtener un estudio comparativo del comportamiento de las diferentes propuestas.

- Estudiar un método, heurística o algoritmo basado en la filosofía del modelo formal para obtener distintos tipos de reglas de asociación.
  - Examinar las distintas propuestas para obtener reglas de asociación y ver cuál se ajusta mejor al modelo.
  - Proponer un algoritmo para la extracción de reglas de asociación siguiendo la filosofía del modelo.

- Estudiar la complejidad del algoritmo y su comportamiento con diversas bases de datos reales.
- Modificar la propuesta para algunos de los tipos de reglas vistas anteriormente: reglas de excepción, anómalas y dobles.

Estructura de la Tesis

La memoria está organizada como se muestra a continuación. En primer lugar tras esta introducción donde se han expuesto los objetivos a cumplir durante el desarrollo del documento, es necesario describir con más precisión el ambiente donde vamos a trabajar y desarrollar las distintas herramientas, además de definir unos conceptos previos y fijar la notación que seguiremos en el resto del trabajo.

Todo lo anterior junto con un estudio sobre los distintos modelos para reglas de asociación y una primera presentación del modelo formal a desarrollar, se presentará en el primer capítulo de Preliminares.

El segundo capítulo engloba todos los Resultados Teóricos que se han desarrollado para cubrir los objetivos anteriormente propuestos. Podemos distinguir cuatro secciones bien diferenciadas. En la Sección 2.1 estudiamos los distintos criterios para medir el interés de las reglas de asociación. Tras su estudio, proponemos y justificamos añadir dos nuevos principios al conjunto de principios ya existentes de Piatetsky-Shapiro en [Piatetsky-Shapiro, 1991] y analizamos la estrecha relación entre los principios y el tipo de cuantificador-4ft del modelo formal bajo estudio. Terminaremos la sección con la definición de nuevos tipos de cuantificadores, estudiando el caso particular del nuevo cuantificador-4ft que definiremos a partir de la medida de interés conocida como Factor de Certeza presentada en [Berzal et al., 2002].

En la Sección 2.2 seguimos con el desarrollo del modelo lógico-formal para la extracción de reglas difusas. Para ello utilizamos el modelo de representación de cantidades imprecisas mediante niveles de restricción [Sánchez et al., 2009] introducido en el Capítulo 1 de preliminares. Gracias a la generalización obtenida, ofrecemos un procedimiento que extiende de forma natural las medidas de interés del caso crisp al difuso.

La Sección 2.3 desarrolla varias propuestas para la extracción de conocimiento en bases de datos formadas por bolsas. Primero definimos los conceptos de bolsa y bolsa difusa. Después continuamos con un análisis de anteriores propuestas y finalizamos con nuestras propuestas. Para el desarrollo de los distintos

métodos, hemos utilizado la teoría de subconjuntos difusos (consultar Apéndice A) y dos importantes herramientas: las reglas difusas y las dependencias graduales [Sánchez et al., 2008d], ésta última presentada recientemente. Mediante el uso de las dependencias graduales conseguimos un tipo de información muy útil puesto que no sólo se muestra la relación entre los items, sino que también nos ofrece una relación entre la variación que sufren las cantidades asociadas.

En la Sección 2.4 se pretende estudiar distintos métodos de extracción de conocimiento mediante reglas de asociación. En particular nos centramos en aquellos que involucran la negación de items y encierran un conocimiento específico que sea de utilidad para el usuario: las reglas de excepción y las reglas anómalas. Las analizamos en profundidad y proponemos nuevas propuestas los dos tipos de reglas. Además usamos el modelo para formalizar las propuestas y compararlas con las ya existentes. Por último proponemos un nuevo tipo de regla a la que denominamos *doble* que junto con las reglas anteriores ofrecen una descripción más completa de la relación existente entre dos conjuntos de items.

El último capítulo muestra un breve repaso a los principales algoritmos de extracción de reglas de asociación deteniéndose en una propuesta que utiliza una representación binaria de los items [Louie and Lin, 2000b] que nos será de utilidad para desarrollar un procedimiento para extraer reglas de asociación utilizando el modelo formal, y en particular, lo utilizaremos para extraer reglas de excepción, anómalas y dobles y comparar las distintas propuestas presentadas en la Sección 2.4.

La memoria terminará con las conclusiones sobre el trabajo realizado y algunas propuestas para futuras líneas a desarrollar en nuestra investigación.

Capítulo 1

Preliminares

Este capítulo contiene los conocimientos previos que deben conocerse para entender el contenido de la tesis. En la sección 1.1 introduciremos un breve resumen sobre los conceptos clave en el campo de la Minería de datos y, en particular, los rudimentos sobre los distintos tipos de reglas de asociación centrándonos en aquellos que sean de utilidad en el siguiente capítulo. La sección 1.2 provee un resumen de algunas de las formulaciones presentadas hasta el momento para la representación y/o evaluación de las reglas de asociación. Por último, en la sección 1.3 se repasan los conceptos básicos de un modelo propuesto en [Sánchez et al., 2008b] para representar y operar con propiedades difusas distinto a la Teoría de Subconjuntos Difusos propuesto por Zadeh.

1.1. Extracción de Conocimiento, Minería de Datos y Reglas de Asociación

Es un hecho que en las últimas décadas el volumen de los datos almacenados ha crecido de forma considerable debido a su guardado en formato digital, a la automatización de muchas tareas, etcétera. El creciente interés por querer utilizar estos datos para obtener información que ayude a aumentar los beneficios en una empresa, a controlar la evolución de los datos, etc, ha favorecido el surgimiento de nuevas herramientas que nos permiten manejar grandes volúmenes de datos y a su vez adquirir información oculta en ellos que pueda ser de utilidad en algún sentido. Este tipo de herramientas son parte del campo de la *Extracción de Conocimiento* (KD, Knowledge Discovery).

De forma abstracta, podríamos decir que la extracción de conocimiento está relacionada con el desarrollo de métodos y técnicas que nos ayuden a entender los datos. Por ejemplo, una de las tareas de las que se ocupa el KD es resumir la información y presentarla de forma inteligible. En el núcleo de este tipo de procesos está la aplicación de algunos métodos específicos de la Minería de Datos (DM).

La denominación *extracción del conocimiento en bases de datos* (KDD) fue acuñada en el primer KDD workshop en 1989. Desde el punto de vista de [Fayyad et al., 1996], KDD es el proceso de descubrir conocimiento útil en los datos, y la minería de datos es la aplicación de algoritmos específicos para extraer *patrones* de los datos, entendiendo el término patrón en un sentido amplio (relaciones, correlaciones, tendencias, agrupamientos, clasificaciones etc.). Procesos como la preparación, la selección y la limpieza de los datos forman parte del proceso de KDD y no de la minería de datos. En la Figura 1.1 se explican de forma clara los pasos de un proceso de KDD [Fayyad et al., 1996].

En [Fayyad et al., 1996] también podemos encontrar una buena aproximación a la definición de extracción del conocimiento:

> *'El proceso para usar una base de datos para cualquier consulta que se requiera; incluyendo: preprocesamiento, muestreo y transformaciones, aplicación de técnicas de minería de datos para obtener patrones y la evaluación de los resultados de dicha minería para identificar qué patrones se consideran conocimiento.'*

Hemos visto que la minería de datos involucra el desarrollo de modelos y algoritmos para determinar patrones de los datos observados, de este modo una

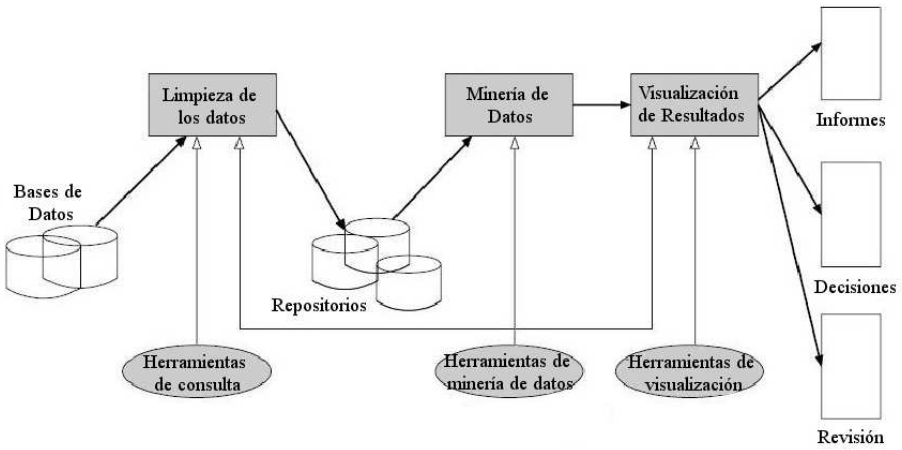

Figura 1.1: Resumen de los pasos que componen un proceso de KDD

de las definiciones más intuitivas de minería de datos puede ser la dada en [Frawley et al., 1992],

> '*la minería de datos es el proceso no trivial de identificar conocimiento en los datos que sea válido, novedoso, potencialmente útil y comprensible*'

Tampoco hay una definición estándar de patrón, aunque Frawley et al. [Frawley et al., 1992], proponen la siguiente:

> '*Dados una serie de hechos (datos) D, un lenguaje L, y una medida de certeza C, un patrón P es un enunciado en L que describe las relaciones (asociaciones) entre varios subconjuntos de D con una certeza dada mediante C, tal que P es más sencillo (en algún sentido) que la enumeración de todas las relaciones entre dichos subconjuntos*'

Esta definición es lo bastante general como para aplicarla a diferentes aproximaciones que producen patrones. Otras propuestas a la definición de patrón pueden encontrarse en [Hand, 2002] y [Höppner, 2005].

A partir de la definición de patrón podríamos considerar, siguiendo las palabras de Frawley et al. [Frawley et al., 1992], que el *conocimiento* lo formarán aquellos patrones que sean *interesantes* (de acuerdo a una medida de interés impuesta por el

usuario) y que se cumplan con certeza suficiente (de acuerdo a un criterio impuesto por el usuario); aunque más adelante veremos que el interés y la certeza no son los dos únicos criterios para obtener conocimiento. Lo que sí debería considerarse es que la extracción del conocimiento debe ser un proceso subjetivo, es decir, se deben tomar en consideración los criterios del usuario.

Este proceso debe hacerse también de forma automática ya que el usuario no puede manipular la enorme cantidad de datos que puede tener a su disposición, y es deseable que dicho conocimiento se exprese mediante una representación apropiada para el entendimiento del usuario.

Algunas representaciones incluyen el lenguaje natural, la lógica formal y representaciones gráficas de la información. El lenguaje natural es a menudo deseable desde el punto de vista humano, pero no lo es en la manipulación de algoritmos de extracción. Las representaciones lógicas son más naturales para la computación y pueden ser traducidas al lenguaje natural. Algunas representaciones lógicas comunes incluyen formalismos para la extracción de reglas (por ejemplo, si X entonces Y), los patrones relacionales ($X > Y$), los árboles de decisión y las redes semánticas o causales. Pero estas formas de manejar el conocimiento tienen tanto ventajas como limitaciones.

Según Fayyad et al. [Fayyad et al., 1996], en la práctica, los dos objetivos principales de la minería de datos son la *predicción* y la *descripción*, aunque hay autores que dividen estos objetivos en dos disciplinas distintas; considerando la predicción como el núcleo principal del *aprendizaje automático* (machine learning), y la descripción una disciplina del dominio de la minería de datos [Hüllermeier, 2008]. Aunque estos dos conceptos no son excluyentes, es necesario distinguirlos y tenerlos en consideración:

- La predicción involucra el uso de varias variables en una base de datos para intentar determinar los valores desconocidos de otras variables, o bien, para adivinar valores futuros que puedan tener las propias variables en la base de datos.

- La descripción se centra en buscar patrones interpretables (por el humano) que expliquen la naturaleza de la base de datos.

Hay diferentes métodos que nos pueden ayudar a conseguir estos dos objetivos. Veremos de forma muy escueta algunos de ellos [Fayyad et al., 1996].

- La *clasificación* tiene como objetivo conseguir pronosticar el valor (o la clase) que puede tomar un atributo concreto en función de los valores que toman otros atributos. Para ello, se han utilizado distintos modelos para representar diferentes tipos de clasificación: conjunto de reglas, listas de decisión, árboles de decisión, etc.
- La *regresión* es una clasificación con consecuente numérico, es decir, pretende pronosticar qué valor numérico tiene un determinado atributo.
- El *agrupamiento* tiene como objetivo conseguir hacer grupos de individuos (tuplas) en función de sus similitudes.
- El *resumen* intenta encontrar una descripción compacta de los datos.
- La *búsqueda* y el *modelado de las dependencias*, consiste en encontrar un modelo que describa las dependencias existentes en los datos.
- La *detección de cambios* se centra en descubrir los cambios a lo largo de un periodo de tiempo en los datos, este objetivo se consigue mediante *modelos de series temporales*.

A continuación detallaremos la descripción de los distintos tipos de reglas de asociación que se han desarrollado hasta el momento para extraer diversos tipos de información. Nos centraremos en aquellas propuestas que vayamos a utilizar en el capítulo 2 o que sirvan para mejorar el entendimiento del resto del documento. También describiremos brevemente algunos de los algoritmos más importantes para la extracción de reglas de asociación.

1.1.1. Reglas de Asociación

Las reglas de asociación son una importante herramienta de la minería de datos que ha recibido mucha atención y estudio desde su primera aparición en los trabajos de Agrawal en [Agrawal et al., 1993]. La idea de las reglas de asociación se originó en el contexto del análisis de bolsas de compra donde se estudia la compra conjunta de varios artículos en una tienda. Las correlaciones entre algunos items[1] particulares, podrían ayudar a organizar algunas estrategias de marketing, promoción de artículos, planificación de los almacenes, etc. En la actualidad el uso de las reglas de asociación se extiende más allá del marketing. Son numerosos y

[1]Sería más correcto hablar de *artículos* en español, pero en lo sucesivo usaremos el término ítem.

notables los distintos campos de aplicación de dichas reglas. Citemos por ejemplo en el ámbito de la Medicina los logros conseguidos al establecer útiles relaciones entre síntomas y diagnósticos [Gamberger et al., 1999], [Ordonez et al., 2000], o en Biología donde el estudio de las relaciones entre genes que actúan conjuntamente en un proceso sigue siendo un campo activo de investigación [López et al., 2007], [Rahal et al., 2008].

En términos generales podríamos decir que las reglas de asociación son expresiones de la forma $X \rightarrow Y$, donde X e Y son conjuntos disjuntos de items, por ejemplo, $\{sobres, papel\} \rightarrow \{sellos\}$ denota la aparición conjunta de los items sobres, papel y sellos en la mayoría de las compras de un estanco. Las reglas de asociación llevan asociadas unas medidas que indican el interés o la confianza que podemos depositar en la asociación, en el ejemplo anterior:

> *el 80 % de los clientes de un estanco que compran sobres y papel, también compran sellos*

nos da una estimación de la probabilidad de que se compren sellos cuando se han comprado sobres y papel. Esta medida es conocida como la confianza de la regla. Otra medida que nos aporta una importante información es el número de compras que avalan la anterior regla, por ejemplo el 10 % total de las compras. Esto es lo que se conoce como soporte.

Formalmente, dados $I = \{i_1, \ldots, i_m\}$ un conjunto finito de items, un *itemsets* será cualquier subconjunto no vacío de I. En este ambiente se denominará *transacción* a cualquier subconjunto no vacío de items y una base de datos D constará de un número determinado de transacciones que notaremos por $|D|$ de ahora en adelante.

El concepto de reglas de asociación puede enunciarse formalmente como sigue [Agrawal et al., 1993]:

Definición 1.1. Sean $I = \{i_1, \ldots, i_m\}$ un conjunto finito de items, D una base de datos donde cada transacción tenga un único identificador y contenga un conjunto de items. Una *regla de asociación* es una implicación de la forma

$$X \longrightarrow Y$$

donde $X, Y \subset I$ son conjuntos de items llamados itemsets cumpliendo que $X \cap Y = \phi$. Llamaremos a X *antecedente* y a Y *consecuente* de la regla de asociación anterior.

En general, X e Y serán dos conjuntos disjuntos de items: $\{x_1, \ldots, x_k\}$ y $\{y_1, \ldots, y_l\}$ respectivamente, donde $x_i, y_j \in I$ para todo $i \in \{1, \ldots, k\}$ y $j \in \{1, \ldots, l\}$.

Observemos que los itemsets podrían representarse en lógica proposicional de forma que $X = x_1 \wedge \ldots \wedge x_k$ lo que nos indica que para que X esté contenida en una transacción, deben estarlo todos los items que la forman. De manera análoga se razonaría con el itemsets Y.

Para evaluar el cumplimiento o grado de verdad de una regla de asociación es habitual utilizar las llamadas *medidas de interés*. Las dos medidas más utilizadas, aunque no por ello las que tienen mejores propiedades, son el soporte y la confianza. Ambas utilizan el concepto de soporte de un itemsets que definimos a continuación.

Definición 1.2. El *soporte* de un itemsets es el porcentaje de transacciones de D que contienen al itemsets, y lo notaremos por $\mathrm{sop}(\cdot)$,

$$\mathrm{sop}(X) = \frac{|\{t \in D \mid X \subseteq t\}|}{|D|}.$$

El soporte de un itemsets es una estimación de la probabilidad de que una transacción en D contenga al itemsets.

En algunas ocasiones, el soporte se utiliza como una medida absoluta que indica el número de transacciones que satisfacen el itemsets. Al utilizar el soporte absoluto hay que tener en cuenta que los cambios producidos en él pueden ser y significar cosas totalmente distintas cuando se producen en un volumen menor o mayor de datos (ver [Brijs et al., 2003]).

Definición 1.3. El *soporte* de una regla de asociación es el porcentaje de transacciones de D que contienen a X y a Y a la vez,

$$\mathrm{Sop}(X) = \frac{|\{t \in D \mid X \cup Y \subseteq t\}|}{|D|}.$$

Definición 1.4. La *confianza* de una regla de asociación es el porcentaje de transacciones de D que contienen a X y a Y de entre todas aquellas que cumplen el antecedente,

$$\mathrm{Conf}(X) = \frac{|\{t \in D \mid X \cup Y \subseteq t\}|}{|\{t \in D \mid X \subseteq t\}|}.$$

El soporte de una regla de asociación s coincide con el soporte del itemsets $X \cup Y$ y la confianza c es una estimación de la probabilidad de Y condicionada a X, es decir,

$$s = \mathrm{sop}(X \rightarrow Y) = \mathrm{sop}(X \cup Y) = P(X \cup Y)$$

y

$$c = \text{Conf}(X \to Y) = P(Y|X) = \frac{P(X \cup Y)}{P(X)} = \frac{\text{sop}(X \cup Y)}{\text{sop}(X)} = \frac{s}{\text{sop}(X)}$$

Un ejemplo de regla de asociación sería

> *El 80 % de los clientes de un estanco que compran sobres, compran también sellos. ($s = 10\,\%$)*

En este ejemplo una transacción podría ser una compra realizada en un estanco, el conjunto de items I sería el conjunto de artículos que se pueden comprar en dicho estanco, $X = \{sobre\}$ y $Y = \{sello\}$. La confianza de la regla $X \to Y$ es del 80 %, aunque a menudo suele representarse también como un valor en el intervalo $[0, 1]$, en este caso sería 0.8. El soporte de esta regla sería el porcentaje de compras realizadas (10 %) en las que se adquirió conjuntamente sobres y sellos.

El problema de la extracción de reglas de asociación BARP (Boolean Association Rule Problem) consiste en encontrar todas aquellas reglas que cumplen los umbrales mínimos de soporte y confianza impuestos por el usuario [Srikant and Agrawal, 1996]. A estas reglas se les llama *reglas fuertes*. Una regla de asociación fuerte es aquella cuyo soporte supera el umbral establecido *minsop*, y cuya confianza supera otro umbral también fijado "a priori" al que llamaremos *minconf*.

Algunos autores [Aggarwal and Yu, 1998] han criticado el uso del soporte y la confianza para definir asociaciones interesantes porque, en general, no es fácil encontrar buenos umbrales para el soporte y la confianza. Estos valores deberían depender del volumen y de la dispersión de los datos, así como del problema particular bajo estudio.

Además, estos umbrales interfieren también en el tiempo de ejecución de los algoritmos que se utilizan para extraer reglas de asociación. Entre los más conocidos se encuentra el algoritmo *Apriori* [Agrawal and Srikant, 1994] que es uno de los algoritmos más básicos para la obtención de reglas de asociación. Este se basa en una interesante y útil propiedad sobre los itemsets frecuentes: si un itemsets X no es frecuente tampoco lo es ningún itemsets que lo contenga y, por supuesto, el contrarrecíproco también es cierto: si X es un itemsets frecuente, también lo es cualquiera de sus subconjuntos.

Esta propiedad da pie a que el BARP pueda ser descompuesto en dos subproblemas más simples [Agrawal and Srikant, 1994]: primero se generan todos los itemsets frecuentes y a continuación se extraen aquellas reglas de asociación

Algoritmo 1.1 : Apriori Básico

Entrada: I, D, $minsop$, $minconf$

Salida: Conjunto de reglas de asociación con soporte y confianza $\geq minsop$ y $minconf$.

Algoritmo:

1. Generar todos los itemsets frecuentes, i.e. con soporte $\geq minsop$.
2. Dado un itemsets frecuente $X = \{x_1, \ldots, x_k\}$ con $k \geq 2$, generar todas las reglas de la forma $X \backslash \{x_j\} \rightarrow \{x_j\}$, siendo el soporte de dicha regla el soporte de X; y la confianza, el cociente de este soporte y el soporte de $X \backslash \{x_j\}$.

que contienen itemsets frecuentes superando los umbrales de soporte y confianza impuestos por el usuario como se muestra en el Algoritmo 1.1.

La obtención de reglas de asociación acarrea algunos problemas en muchos casos. Para evitar algunos de ellos se han propuesto otros tipos de reglas alternativas. En las siguientes secciones se estudiarán en detalle las propuestas que se utilizarán en posteriores secciones del documento.

1.1.2. Reglas de Asociación Difusas

Las técnicas de minería de datos y en particular las reglas de asociación, pretenden identificar conocimiento que sea válido, novedoso, útil y *comprensible*. Esto justifica la importancia de trabajar con representaciones del conocimiento que sean *semánticamente significativas* para el usuario. El usuario es la pieza clave en la búsqueda de conocimiento. Este debe guiar al sistema y juzgar la calidad de las relaciones obtenidas para determinar si el conocimiento obtenido es *interesante* o no. Una de las técnicas más utilizadas para representar posibles tipos de imprecisión y vaguedad en la información es la Teoría de Subconjuntos Difusos propuesta por L.A. Zadeh [Zadeh, 1965] cuyos conceptos más importantes se detallan en el Apéndice A. Esta teoría permite utilizar de forma bastante intuitiva distintos tipos de incertidumbre e imprecisión en los datos y extiende las operaciones entre conjuntos usuales para su utilización con conjuntos difusos, permitiendo en nuestro caso extender el concepto de regla de asociación al caso difuso. Cuando pueda haber confusión o queramos distinguir ambos tipos de reglas de asociación utilizaremos los términos *crisp* y *difuso*. La teoría de Subconjuntos Difusos es muy útil en muchos aspectos para la extracción de reglas de asociación difusas, pero adolece de una buena extensión de la negación de conjuntos que

cumpla algunas propiedades lógicas entre conjuntos. Por este motivo se han propuesto otros modelos para representar imprecisión en los datos como por ejemplo la Representación mediante niveles de restricción [Sánchez et al., 2008b] que soluciona estos aspectos y que también utilizaremos en la sección 1.3.

Los primeros trabajos que empezaron a hablar de reglas de asociación difusas se centraban en el uso de etiquetas difusas para conseguir reglas de asociación cuantitativas [Hong and Lee, 2008], [Delgado et al., 2003b]. En las reglas de asociación cuantitativas se divide el dominio del atributo en intervalos para después extraer reglas de asociación cuyos items son pares del tipo $\langle atributo, intervalo\rangle$ en lugar de $\langle atributo, valor\rangle$. Por ejemplo si tenemos el conjunto de reglas

$$\langle \text{altura} \ , 1.71 \ m\rangle \rightarrow \langle \text{peso} \ , 81 \ kg\rangle$$
$$\langle \text{altura} \ , 1.75 \ m\rangle \rightarrow \langle \text{peso} \ , 90 \ kg\rangle$$
$$\langle \text{altura} \ , 1.80 \ m\rangle \rightarrow \langle \text{peso} \ , 85 \ kg\rangle$$

podríamos englobarlas en una sola regla del tipo

$$\langle \text{altura}, (1.70, 1, 80]\rangle \rightarrow \langle \text{peso}, (80, 90]\rangle \, .$$

Pero al dividir el dominio en un número fijo de intervalos surge el llamado *problema de la frontera* en el que podemos rechazar algunos intervalos interesantes al excluir elementos potenciales cerca de su frontera (ver [Kuok et al., 1998]). Para solucionar este problema algunos autores han propuesto dividir el dominio del atributo en intervalos solapados. Pero con la Teoría de subconjuntos difusos, podemos evitar el problema de definir cuáles serán los bordes del intervalo utilizando en su lugar conjuntos difusos. De esta forma un elemento pertenecerá a un conjunto difuso con un grado de pertenencia en $[0, 1]$ y además podremos dotar a dichos conjuntos de una semántica significativa para el entendimiento del usuario utilizando para ello términos lingüísticos. En el ejemplo anterior tendríamos una regla de asociación difusa del tipo

$$\langle \text{altura}, alto\rangle \rightarrow \langle \text{peso}, bastante \ pesado\rangle \, .$$

En esta línea se encuentran los primeros trabajos [Lee and Kwang, 1997], [Au and Chan, 1997], [Kuok et al., 1998] y [Gyenesei, 2001] que datan de finales de los años 90. Otros trabajan con un enfoque distinto en el que las reglas de asociación difusas fueron introducidas para manejar reglas cuyos items tienen asociado un grado difuso que recoge su importancia relativa en la regla [Yue et al., 2000], [Cai et al., 1998]. Este enfoque difiere de los anteriores en que no utiliza etiquetas lingüísticas.

A continuación desarrollaremos las dos propuestas más representativas para reglas de asociación difusas. La primera, es la elaborada por Kuok, Gyenesey y otros autores [Gyenesei, 2001], [Kuok et al., 1998] basada en la extracción de reglas de asociación difusas en bases de datos relacionales crisp. La segunda es la realizada por Delgado et al. en [Delgado et al., 2003a] en la que aparecen los conceptos de transacción y regla difusa además de un nuevo enfoque para evaluar reglas de asociación difusas usando sentencias cuantificadas (ver Apéndice A).

Enfoque de Kuok & Gyenesei

Consideremos $D = \{t_1, \ldots, t_n\}$ una base de datos transaccional con atributos I y los conjuntos difusos asociados a dichos atributos. Para Gyenesei et al. [Gyenesei, 2001] una *regla de asociación difusa* es una expresión de la forma:

$$\begin{aligned} &\text{si } X = \{x_1, \ldots, x_p\} \text{ es } A = \{a_1, \ldots, a_p\} \text{ entonces} \\ &\quad Y = \{y_1, \ldots, y_q\} \text{ es } B = \{b_1, \ldots, b_q\}, \end{aligned}$$

donde

$$\begin{aligned} a_i &\in \{\text{conjuntos difusos asociados a } x_i\} \\ b_i &\in \{\text{conjuntos difusos asociados a } y_i\}, \end{aligned}$$

X, Y son conjuntos de items de I disjuntos, es decir, no tienen atributos comunes, y A, B contienen los conjuntos difusos asociados a los correspondientes atributos de X e Y respectivamente.

A la primera parte de la regla 'X es A' se le llama antecedente y la segunda, 'Y es B' es el consecuente de la regla. El significado de la regla es que cuando se *satisface* que X es A entonces Y es B también se cumple. Aquí satisfacer significa que hay un número suficiente de transacciones que contribuyen a superar un umbral mínimo especificado por el usuario. De ahora en adelante notaremos dicha regla por $\langle X, A\rangle \rightarrow \langle Y, B\rangle$.

Para calcular el soporte difuso de un itemsets del tipo $\langle X, A\rangle$ proponen utilizar la fórmula siguiente:

$$\text{FSop}(\langle X, A\rangle) = \frac{\sum_{t_i \in D} \prod_{x_j \in X} \mu^i_{a_j}(x_j)}{|D|} \tag{1.1}$$

donde $|D|$ es el número de transacciones de la base de datos y $\mu^i_{a_j}(x_j)$ es el grado de pertenencia del atributo $x_j \in X$ en la transacción i-ésima al conjunto difuso $a_j \in A$.

$\langle edad, media\rangle$	$\langle sueldo, bajo\rangle$
0.7	0.5
0.2	0.3
0.5	0.2
0.3	0.4
0.6	0.2
0.8	0.4

Tabla 1.1: Parte de una Base de Datos que contiene grados de pertenencia de los items $\langle edad, media\rangle$ y $\langle sueldo, bajo\rangle$.

Observemos que en la fórmula (1.1) se podría haber utilizado en lugar del producto cualquier otra t-norma. De forma análoga al caso crisp, para calcular el soporte y la confianza de la regla de asociación difusa $\langle X, A\rangle \rightarrow \langle Y, B\rangle$, usaremos las fórmulas:

$$\begin{aligned} \mathrm{FSop}(\langle X, A\rangle \rightarrow \langle Y, B\rangle) &= \mathrm{FSop}(\langle Z, C\rangle) \\ \mathrm{FConf}(\langle X, A\rangle \rightarrow \langle Y, B\rangle) &= \frac{\mathrm{FSop}(\langle Z, C\rangle)}{\mathrm{FSop}(\langle X, A\rangle)} \end{aligned} \tag{1.2}$$

donde $Z = X \cup Y$ y $C = A \cup B$.

Para entender mejor el cálculo del soporte y la confianza presentaremos un ejemplo. Supongamos que una base de datos D tiene a *edad* y *sueldo* entre sus atributos con conjuntos difusos asociados {niño, joven, mayor, anciano} y {alto, medio, bajo} respectivamente. Una regla de asociación que podría encontrarse en D sería $\langle edad, media\rangle \rightarrow \langle sueldo, bajo\rangle$ con soporte y confianza calculados a continuación tomando los valores dados en la Tabla 1.1.

$$\begin{aligned} \mathrm{FSop}(\langle edad, media\rangle \rightarrow \langle salario, bajo\rangle) &= \mathrm{FSop}(\langle \{edad, salario\}, \{media, bajo\}\rangle) \\ &= \frac{0.35 + 0.06 + 0.1 + 0.12 + 0.12 + 0.32}{6} \\ &= 0.178 \end{aligned}$$

$$\begin{aligned} \mathrm{FConf}(\langle edad, media\rangle \rightarrow \langle salario, bajo\rangle) &= \frac{0.35 + 0.06 + 0.1 + 0.12 + 0.12 + 0.32}{0.7 + 0.2 + 0.5 + 0.3 + 0.6 + 0.8} \\ &= 0.345 \end{aligned}$$

Además del soporte y la confianza, otras medidas del cumplimiento para reglas de asociación han sido tenidas en cuenta para su extensión al caso difuso [Gyenesei, 2001], [Kuok et al., 1998].

Enfoque de Delgado et al.

En los trabajos [Delgado et al., 2003a] y [Delgado et al., 2003b] se generaliza el concepto de transacción y regla de asociación para el caso en el que estas sean difusas y también se introduce una forma de medir el soporte y la validez de las reglas difusas que generaliza de forma natural el caso crisp y engloba al enfoque anterior. Veremos los conceptos más importantes a continuación.

Definición 1.5. [Delgado et al., 2003a] Sea $I = \{i_1, \ldots, i_m\}$ un conjunto finito de items. Una transacción difusa es un subconjunto difuso no vacío $\tilde{\tau} \subseteq I$.

Para cada item $i \in I$ y cada transacción $\tilde{\tau}$ tendremos que i pertenece a $\tilde{\tau}$ con grado[2] $\tilde{\tau}(i)$, donde $\tilde{\tau}(i)$ es un número real en el intervalo $[0, 1]$.

Sea A un itemsets de I, es decir, un subconjunto de items, en una transacción difusa $\tilde{\tau}$, se define el grado de pertenencia de $A \subseteq I$ a la transacción difusa $\tilde{\tau}$ como

$$\tilde{\tau}(A) = \min_{i \in A} \tilde{\tau}(i). \tag{1.3}$$

Usando la definición 1.5 una transacción será un caso especial de transacción difusa donde todos los items tienen valor 0 ó 1 según se encuentren o no en la transacción.

Ejemplo 1.1. Consideremos el conjunto de items $I = \{i_1, i_2, i_3, i_4\}$ y la base de datos transaccional difusa dada por la Tabla 1.2.

	i_1	i_2	i_3	i_4
$\tilde{\tau}_1$	0	0.6	0.7	0.9
$\tilde{\tau}_2$	0	1	0	1
$\tilde{\tau}_3$	1	0.5	0.75	1
$\tilde{\tau}_4$	1	0	0.1	1
$\tilde{\tau}_5$	0.5	1	0	1
$\tilde{\tau}_6$	1	0	0.75	1

Tabla 1.2: Base de datos transaccional difusa.

En particular, podemos observar que $\tilde{\tau}_2$ es una transacción crisp. Algunos grados de inclusión son: $\tilde{\tau}_1(\{i_3, i_4\}) = 0.7$, $\tilde{\tau}_1(\{i_2, i_3, i_4\}) = 0.6$ y $\tilde{\tau}_4(\{i_1, i_4\}) = 1$. ♦

[2] Observemos que $\tilde{\tau}(i)$ no es más que $\mu_{\tilde{\tau}}(i)$ donde $\mu_{\tilde{\tau}} : I \longrightarrow [0, 1]$ es la función de pertenencia del conjunto difuso $\tilde{\tau}$ sobre el conjunto de items I.

Definición 1.6. [Delgado et al., 2003a] Sea I un conjunto de items, D un conjunto de transacciones difusas y $A, B \in I$ dos itemsets disjuntos, es decir, $A \cap B = \phi$. Una regla de asociación difusa $A \rightarrow B$ se cumple en D si y sólo si,

$$\tilde{\tau}(A) \leq \tilde{\tau}(B) \tag{1.4}$$

para todo $\tilde{\tau} \in D$, esto es, el grado de inclusión de B es mayor que el grado de inclusión de A para todas las transacciones difusas $\tilde{\tau}$ en D.

Esta definición preserva el significado de las reglas de asociación crisp ya que si necesitamos que $A \subseteq \tilde{\tau}$ en algún sentido, necesitaremos que también $B \subseteq \tilde{\tau}$ se cumpla, lo que en nuestro caso se traduce en la desigualdad $\tilde{\tau}(A) \leq \tilde{\tau}(B)$. De esta forma, como una transacción crisp es un caso especial de transacción difusa, una regla de asociación será un caso especial de regla de asociación difusa.

Usando este modelo en [Delgado et al., 2003a] y [Delgado et al., 2003b] se presenta una generalización de las medidas de soporte y confianza usando una aproximación semántica basada en la evaluación de sentencias cuantificadas.

La evaluación de una sentencia cuantificada es un valor en el intervalo $[0, 1]$ que nos dice el grado de cumplimiento de una expresión del tipo "Q de los F son G", donde F y G son dos subconjuntos de un conjunto finito X, y Q es un cuantificador relativo difuso. Ejemplos de cuantificadores relativos son "la mayoría", "muchos", o "casi todos" y vienen definidos por conjuntos difusos (ver Figura 1.2). Para calcular el grado de cumplimiento de una sentencia cuantificada existen varios métodos que podemos encontrar con más detalle en el Apéndice A. En particular, en [Delgado et al., 2003a] se utiliza el método GD propuesto en [Delgado et al., 2000] y es el que detallamos en las ecuaciones (1.7) y (1.9).

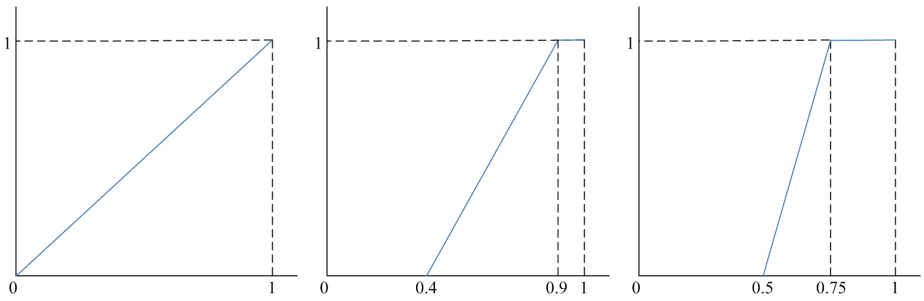

Figura 1.2: Ejemplos de cuantificadores relativos difusos. De izq a dcha: "la mayoría", "muchos" y "casi todos".

Definición 1.7. [Delgado et al., 2003a] Sea $A \subseteq I$. El *soporte* de A en D es la evaluación de la sentencia cuantificada

$$Q \text{ de los } D \text{ son } \tilde{\Gamma}_A \tag{1.5}$$

donde $\tilde{\Gamma}_A$ es un conjunto difuso en D definido como $\tilde{\Gamma}_A(\tilde{\tau}) = \tilde{\tau}(A)$.

Definición 1.8. [Delgado et al., 2003a] El *soporte* de la regla de asociación difusa $A \rightarrow B$ en el conjunto de transacciones difusas D es $Sop(A \cup B)$, es decir, la evaluación de la sentencia cuantificada

$$Q \text{ de los } D \text{ son } \tilde{\Gamma}_{A\cup B}. \tag{1.6}$$

Usando el método GD y el cuantificador $Q =$ la mayoría, el soporte de una regla difusa $A \rightarrow B$ es:

$$\begin{aligned} \text{Sop}(A \rightarrow B) &= GD_{Q_M}\left(\tilde{\Gamma}_{A\cup B}/D\right) \\ &= \sum_{\alpha_i \in \Delta(\tilde{\Gamma}_{A\cup B}/D)} (\alpha_i - \alpha_{i+1}) Q_M \left(\frac{\left|(\tilde{\Gamma}_{A\cup B} \cap D)_{\alpha_i}\right|}{|D_{\alpha_i}|} \right) \end{aligned} \tag{1.7}$$

donde $Q_M(x) = x$ es el cuantificador relativo 'la mayoría', $\Delta(\tilde{\Gamma}_{A\cup B}/D) = \Lambda(\tilde{\Gamma}_{A\cup B}) \cup \Lambda(D)$, y $\Delta(\tilde{\Gamma}_{A\cup B}/D) = \{\alpha_1, \ldots, \alpha_p\}$ con $\alpha_i > \alpha_{i+1}$ para cualquier $i \in \{1, \ldots, p\}$.

Definición 1.9. [Delgado et al., 2003a] La *confianza* de la regla de asociación difusa $A \rightarrow B$ en el conjunto de transacciones difusas D es la evaluación de la sentencia cuantificada

$$Q \text{ de los } \tilde{\Gamma}_A \text{ son } \tilde{\Gamma}_B. \tag{1.8}$$

De manera similar al soporte y usando de nuevo el cuantificador relativo $Q_M(x) = x$, definimos la confianza de la regla difusa $A \rightarrow B$ como

$$\begin{aligned} \text{Conf}(A \rightarrow B) &= GD_{Q_M}\left(\tilde{\Gamma}_B/\tilde{\Gamma}_A\right) = \\ &= \sum_{\alpha_i \in\ \Delta(\tilde{\Gamma}_B/\tilde{\Gamma}_A)} (\alpha_i - \alpha_{i+1}) Q_M \left(\frac{\left|(\tilde{\Gamma}_A \cap \tilde{\Gamma}_B)_{\alpha_i}\right|}{\left|(\tilde{\Gamma}_A)_{\alpha_i}\right|} \right) \end{aligned} \tag{1.9}$$

donde $\Delta(\tilde{\Gamma}_B/\tilde{\Gamma}_A) = \Lambda(\tilde{\Gamma}_A \cap \tilde{\Gamma}_B) \cup \Lambda(\tilde{\Gamma}_A)$, y $\Delta(\tilde{\Gamma}_B/\tilde{\Gamma}_A) = \{\alpha_1, \ldots, \alpha_p\}$ con $\alpha_i > \alpha_{i+1}$ para cualquier $i \in \{1, \ldots, p\}$. El conjunto $\tilde{\Gamma}_A$ debe estar normalizado, si no, se normalizaría y el mismo factor de normalización se aplicaría a la intersección $\tilde{\Gamma}_B \cap \tilde{\Gamma}_A$.

Conviene señalar que estas definiciones establecen familias de medidas de soporte y confianza, dependiendo del método de evaluación y del cuantificador Q elegido. En los trabajos [Delgado et al., 2003a] y [Delgado et al., 2003b] se justifica el uso del método GD para evaluar las sentencias y el cuantificador $Q = Q_M$ donde $Q_M(x) = x$.

Otros enfoques

Los anteriores enfoques son las dos alternativas más representativas para evaluar y extraer reglas de asociación difusas, aunque en la literatura se pueden encontrar otros modelos y diversos algoritmos para obtener este tipo de reglas. En [Delgado et al., 2003b] y [Hong and Lee, 2008] podemos encontrar resumidas las técnicas más relevantes en este área. También merecen mención algunos trabajos [Sudkamp, 2005], [Cock et al., 2005], [Dubois et al., 2006] que tratan la generalización de las medidas de interés para reglas difusas tema que trataremos con mayor profundidad en el próximo capítulo.

Aunque la utilización de reglas difusas está debidamente motivada, también queremos destacar su uso para desarrollar diversas aplicaciones en distintos campos de la ciencia [López et al., 2007], [Sánchez et al., 2008e] y en otras áreas como por ejemplo en Aprendizaje Automático [Hüllermeier, 2008]. En problemas de clasificación se han utilizado con muy buenos resultados y también han servido como herramienta para extraer otros tipos de patrones: patrones secuenciales [Fiot et al., 2008a], dependencias difusas [Molina et al., 2008b], etc.

1.1.3. Descubrimiento de Dependencias usando Reglas de Asociación

Las reglas de asociación se han usado en diversos tipos de estructuras para obtener distintos tipos de información definiendo para ello tipos especiales de items y/o transacciones. Un caso especial son las *dependencias funcionales* o las *dependencias graduales*.

Sea RE un conjunto de atributos y r una instancia de RE. Una dependencia funcional $X \rightarrow Y$, $X, Y \subset ER$, se cumple en r si el valor de $t[X]$ determina el de $t[Y]$ para cada tupla $t \in r$. La dependencia se cumplirá en RE si se satisface en cada instancia de RE. Formalmente, una dependencia funcional se satisface en una relación r si, y sólo si,

$$\forall\, t, s \in r \quad \text{si} \;\; t[X] = s[X] \;\; \text{implica que} \;\; t[Y] = s[Y]. \tag{1.10}$$

Encontrar dependencias funcionales es muy interesante, pero cuando utilizamos reglas de asociación no es fácil encontrar este tipo de dependencias debido a que es muy difícil que todas las instancias satisfagan la condición anterior. Para suavizar esta definición se han propuesto principalmente dos grupos diferentes de dependencias: las *dependencias funcionales difusas* y las *dependencias aproximadas*. La primera, introduce algunas componentes difusas (e.g. la igualdad se puede reemplazar por una relación de similitud); mientras, la segunda introduce las dependencias funcionales con un grado de incertidumbre. Las dependencias aproximadas también se pueden ver como dependencias funcionales en las que se ha relajado el cuantificador universal $\forall$. En [Molina et al., 2008b] podemos encontrar un resumen de los diferentes tipos de dependencias así como nuevas propuestas de dependencias difusas.

Un tipo completamente diferente de dependencias son las dependencias graduales presentadas en [Hüllermeier, 2002]. Las llamadas *dependencias graduales* se definen para medir una variación en la relación existente entre dos tipos de objetos. Estas dependencias difieren de las anteriores en que permiten expresar una "tendencia" en los datos en lugar de ver solamente cuándo se satisface la igualdad (o la relación de similitud, en el caso difuso). Un ejemplo de dependencia gradual es la regla *cuanto mayor el salario, mejor el puesto de trabajo*, significando que cuando el salario de una persona crece, esta posee un mayor rango en su trabajo. Este tipo de regla puede encontrarse por ejemplo en una base de datos de una compañía formada por sus empleados, sus puestos en la compañía y sus salarios.

Los cambios en el grado de pertenencia considerados en las dependencias graduales pueden ser de dos tipos: *mayor* o *menor*. Luego podremos considerar cuatro tipos diferentes de dependencias graduales: *cuanto mayor* X *es* A*, mayor* Y *es* B (expresado como $(>, X, A) \rightarrow (>, Y, B)$), *cuanto mayor* X *es* A*, menor* Y *es* B expresado como $(>, X, A) \rightarrow (<, Y, B)$), etc.

El primer trabajo en el que se habla del uso de reglas de asociación para medir dependencias graduales [Hüllermeier, 2002] basa la evaluación de las reglas en el análisis de la regresión lineal entre las variables. El punto de partida es un *diagrama de contingencia*. Un diagrama de contingencia es una forma de expresar los itemsets difusos del tipo $\langle X, A\rangle$, $\langle Y, B\rangle$ como puntos en un plano, es decir, si X, Y son dos atributos y A, B son conjuntos difusos definidos en X y en Y respectivamente, entonces $(A(x), B(y))$ es un punto de dicho diagrama de contingencia con $x \in X$ e $y \in Y$. Una dependencia gradual vendrá dada por una regla de tendencia del tipo $A \rightharpoonup^t B$, significando que "un incremento en $A(x)$ viene dado con un incremento en $B(y)$". La evaluación de la regla se basa en los coeficientes de regresión $[\alpha, \beta]$

de la recta que aproxima a los puntos $(A(x), B(y))$ del diagrama de contingencia. α será la pendiente de dicha recta y la calidad de la recta de regresión vendrá dada por el coeficiente de correlación usual R^2.

Según el tipo de tendencia que queramos reconocer en los datos así como dependiendo de la naturaleza de los mismos, podemos usar dependencias graduales crisp o difusas. En los siguientes apartados veremos cómo se extraen las dependencias graduales crisp [Berzal et al., 2007] y el mecanismo para pasar al caso difuso mediante la definición del *grado de variación* de un itemsets difuso [Molina et al., 2007], [Molina et al., 2008a].

Dependencias Graduales Crisp

Además de la propuesta hecha en [Hüllermeier, 2002], queremos destacar la alternativa propuesta en [Berzal et al., 2007] para la extracción de dependencias graduales usando reglas de asociación. En el siguiente capítulo utilizaremos esta alternativa para obtener tendencias muy interesantes en un tipo especial de base de datos llamadas *bolsas* o *multisets*.

Una dependencia gradual [Berzal et al., 2007], como la utilizaremos de ahora en adelante, es una regla de la forma $(\rhd_1, X, A) \rightarrow (\rhd_2, Y, B)$, con $\rhd_1, \rhd_2 \in \{<, >\}$. La dependencia se cumplirá en una base de datos D si y sólo si para cualesquiera $(x, y), (x', y') \in D$, el cumplimiento de $A(x) \rhd_1 A(x')$ implica que $B(y) \rhd_2 B(y')$. Para extraer este tipo de dependencias usando reglas de asociación, tendremos que hacerlo en un conjunto apropiado de transacciones de D.

En este contexto en [Berzal et al., 2007] se define el conjunto de items GI^D con items de la forma $[>, X, A], [<, X, A], [>, Y, B], [<, Y, B]$ que expresarán las dos posibles tendencias de los atributos X, Y con respecto a las restricciones marcadas por los conjuntos difusos A y B respectivamente. También se define el conjunto de transacciones, que notaremos por GT^D, como aquel que contiene los items de GI^D obtenidos de D como sigue: para cada par $o = (x, y)$, $o' = (x', y') \in D$ existe una transacción $gt_{oo'} \in GT^D$ tal que $[\rhd, X, A]$ pertenecerá a $gt_{oo'}$ si, y sólo si, $A(x) \rhd A(x')$ con $\rhd \in \{<, >\}$ (de forma similar para $[\rhd, Y, B]$). Entonces, diremos que la dependencia gradual $(\rhd_1, X, A) \rightarrow (\rhd_2, Y, B)$ se cumple en D si, y sólo si la regla de asociación (crisp) $[\rhd_1, X, A] \rightarrow [\rhd_2, Y, B]$ se cumple en GT^D. Las medidas de soporte y confianza definidas para la extracción de reglas de asociación podrán ser utilizadas para validar la anterior dependencia gradual.

La ventaja de este método es que los algoritmos desarrollados para la obtención de reglas de asociación pueden ser fácilmente modificados para extraer

dependencias graduales. Sin embargo, la semántica de esta propuesta y la vista anteriormente en [Hüllermeier, 2002] es que en esta se considera la magnitud de la variación de ambas variables (ver [Molina et al., 2008a]). Para incorporar la magnitud de la variación del grado de cumplimiento de la dependencia gradual, en [Molina et al., 2007] se presenta una nueva propuesta basándose en el concepto de regla de asociación difusa. Haremos un breve resumen de la propuesta a continuación.

Dependencias Graduales Difusas

En la anterior propuesta sólo se tiene en cuenta si el grado de pertenencia es mayor o menor. Así, el grado de pertenencia del item $[\rhd, X, A]$ en una transacción $gt_{oo'} \in GT^D$ puede definirse por la ecuación (1.11)

$$gt_{oo'}([\rhd, X, A]) = \begin{cases} 1 & \text{si } A(x) \rhd A(x') \\ 0 & \text{en otro caso.} \end{cases} \tag{1.11}$$

Con esta definición, para $A(x) = 0$ y $A(x') = 0.1$ obtenemos el mismo resultado que para $A(x) = 0$ y $A(x') = 1$. Esto nos puede llevar a obtener el mismo valor de cumplimiento para dependencias que son intuitivamente diferentes (ver [Molina et al., 2007] para una exposición más detallada del problema). Para evitar este problema, en [Molina et al., 2008a] se propone reemplazar la ecuación (1.11) por una expresión que de un grado en el intervalo $[0, 1]$. Este grado se llamará *grado de variación*. Existen varias posibilidades para obtener dicho grado, pero usaremos el propuesto en [Molina et al., 2008a]:

$$gt_{oo'}([\rhd, X, A]) = \begin{cases} |A(x) - A(x')| & \text{si } A(x) \rhd A(x') \\ 0 & \text{en otro caso.} \end{cases} \tag{1.12}$$

Con esta definición la función $gt_{oo'}$ que mide el grado de variación (1.12) verifica las siguientes propiedades:

(i) $gt_{oo'}([\rhd, X, A]) \in [0, 1]$.

(ii) Si $|A(x) - A(x')| > |A(x) - A(x'')|$ entonces $gt_{oo'}([\rhd, X, A]) > gt_{oo''}([\rhd, X, A])$.

(iii) $gt_{oo'}([<, X, A]) = gt_{o'o}([>, X, A])$.

Teniendo en cuenta el grado de variación, usaremos la definición de dependencias graduales dado en [Molina et al., 2008a].

Sea $GI^D = \{[>, X, A], [<, X, A], [>, Y, B], [<, Y, B]\}$ un conjunto de items y GT^D un conjunto de transacciones difusas que contienen los items de GI^D. GT^D se obtiene de D como sigue: para cualesquiera $o = (x, y)$, $o' = (x', y') \in D$ existe una transacción difusa $gt_{oo'} \in GT^D$ tal que $gt_{oo'}$ es la función dada en la ecuación (1.12), con $\rhd \in \{<, >\}$. Diremos que una regla difusa $[\rhd_1, X, A] \rightarrow [\rhd_2, Y, B]$ en GT^D define una *dependencia gradual difusa* $(\rhd_1, X, A) \rightarrow (\rhd_2, Y, B)$ en D, i.e., las reglas difusas obtenidas en GT^D definen un tipo particular de dependencia gradual difusa en D. En resumen, podremos emplear las medidas de soporte y confianza vistas en la sección 1.1.2 (u otras medidas de interés) para reglas difusas para medir la validez o el interés de un tipo particular de dependencia gradual difusa.

1.1.4. Últimas Tendencias

Las secciones anteriores muestran el potencial que poseen las reglas de asociación y cómo esta herramienta ha ido evolucionando para dar lugar a asociaciones más complejas que permiten extraer información muy interesante para el usuario. Queremos destacar que lo presentado anteriormente es sólo una pequeña muestra de la amplia capacidad de esta herramienta para amoldarse a la extracción de distintos tipos de conocimiento. En este sentido, es obligado mencionar algunas de las áreas que están siendo desarrolladas en la actualidad. Para tener más detalle se pueden consultar [Hipp et al., 2000], [Dunham et al., 2000], [Goethals, 2003], [Delgado et al., 2003b], [Geng and Hamilton, 2006], [Hong and Lee, 2008] y sus referencias.

En la actualidad, se siguen desarrollando algoritmos que sean eficientes para extraer reglas de asociación en bases de datos de una gran extensión [Farzanyar et al., 2006]. También se están estudiando cómo conseguir un conjunto de reglas que sea a la vez mínimo, representativo y que contenga la mayor información posible [Zaki, 2004], [Zaki and Hsiao, 2005]. Otras de las líneas de reciente interés son los nuevos tipos de conocimiento que pueden ser extraídos usando como herramienta las reglas de asociación o modificándolas adecuadamente. Prueba de ello son las reglas de asociación difusas y las dependencias que hemos presentado anteriormente. Además se pueden encontrar otros muchos ejemplos: reglas secuenciales y temporales [Roddick and Spiliopoulou, 2002], [Chen and Huang, 2006], [Fiot et al., 2008b], reglas maximales, reglas con múltiples soportes [Kiran and Reddy, 2009], reglas multimedia, reglas de excepción [Duval et al., 2007], etcétera.

Otras investigaciones en esta área se han enfocado en el estudio de las medidas de interés para la extracción de las reglas de asociación. Pueden destacarse dos importantes tipos de medidas: las objetivas, en las que la participación del usuario se ve delegado a la imposición de los umbrales que deben satisfacer dichas medidas; y las subjetivas, donde el usuario es parte activa del proceso. Esta línea de investigación está siendo desarrollada mediante otras herramientas que permitan al usuario expresar su conocimiento, y a partir de este extraer las reglas de asociación que le resulten útiles, novedosas y lo más interesantes posible [Liu et al., 2000], [Sahar, 2002], [Sinoara and Rezende, 2006], objetivo que pretende conseguir la minería de datos.

A continuación presentaremos el modelo formal que utilizaremos para representar y evaluar las reglas de asociación, analizando sus conceptos y propiedades más importantes. También repasaremos algunas de las propuestas más interesantes usadas para modelizar, representar o analizar propiedades de interés tanto en reglas de asociación como en conjuntos de ellas.

1.2. Modelo Formal para el Estudio de Reglas de Asociación

Desde la aparición de las reglas de asociación, han sido numerosas las aportaciones que han ayudado al desarrollo de distintas facetas en las que interviene esta útil herramienta. Así podemos encontrar distintos tipos de algoritmos para su extracción, nuevas propuestas de medidas de interés, nuevas aplicaciones, nuevos tipos de reglas que utilizan como base las reglas de asociación, etc. pero ha habido pocos esfuerzos en desarrollar un modelo formal que recoja tanto el significado de las reglas de asociación como el tipo de medida para estimar su validez, utilidad o alguna otra característica de interés.

El principal problema para obtener un modelo formal que reúna todas las características deseables para un buen estudio de las reglas de asociación es fijar la propiedad que se quiere analizar, estudiar o desarrollar. Así, podemos encontrarnos con diversas formulaciones que se centran en aspectos concretos de las reglas de asociación pero que no llegan a ofrecer una visión global o unificada sobre dos de sus aspectos más importantes: su representación y su evaluación.

Los modelos que presentaremos a continuación están entre los pocos avances que se han hecho para modelizar la extracción o el comportamiento de las reglas de asociación. Las propiedades inherentes de cada uno han facilitado, según el caso, la extracción de un conjunto mínimo de reglas con la mayor información posible y la extensión de medidas de interés para validar las reglas de asociación difusas. En el resto de la sección, presentaremos el modelo formal que usaremos y desarrollaremos en la tesis, elegido para este propósito por varias razones:

1. El modelo posee buenas propiedades para modelizar las reglas de asociación así como las medidas de interés usadas para su extracción.

2. El modelo se adapta fácilmente a la extracción de distintos tipos de reglas, dando como resultado nuevas medidas de interés que se ajustan a cada tipo de regla.

3. Para ofrecer un marco unificado para manejar las reglas de asociación junto a las medidas de interés utilizadas en su extracción. Esto nos será útil para implementar un algoritmo que sirva para extraer cualquier tipo de regla sin más que cambiar el tipo de medida de interés establecida en términos del modelo.

1.2.1. Algunos modelos en la literatura

A continuación presentaremos algunas de las formulaciones propuestas para la extracción tanto de reglas de asociación, como de conjuntos de reglas que tienen una característica común; presentando sus características más notables y el motivo por las que fueron desarrolladas.

Representación mediante el Retículo de Itemsets Frecuentes y Cerrados

El modelo se basa en la teoría de retículos de Galois [Davey and Priestley, 1994] y permite representar mediante un retículo los itemsets frecuentes y cerrados (ver definición 1.14). Este retículo contiene la información suficiente para obtener un conjunto de reglas minimal en el sentido de que el resto de reglas de asociación pueden obtenerse a partir de ellas. Basándose en este tipo de estructura se han desarrollado diversos algoritmos que se encargan de generar el retículo de los itemsets cerrados para la obtención del conjunto mínimo de reglas de asociación que contengan la mayor información posible. Entre estos algoritmos podemos destacar los siguientes: A-Close [Pasquier et al., 1998], CLOSET [Pei et al., 2000], FCIL (Frequent Closed Itemsets Lattice algorithm) [Jia et al., 2003], CloseMiner [Singh et al., 2005] y CHARM-L [Zaki and Hsiao, 2005].

Las siguientes definiciones pueden encontrarse en [Davey and Priestley, 1994] referidas a los conceptos sobre la Teoría de Retículos de Galois. Podemos encontrarlas tal y como se presentan aquí en [Pasquier et al., 1998].

Definición 1.10. Un *contexto de Minería de Datos* es una terna $\mathcal{D} = (\Gamma, \mathcal{T}, \mathcal{R})$, donde Γ representa el conjunto de transacciones (filas), $\mathcal{T}$ son los itemsets (columnas), y $\mathcal{R} \subseteq \Gamma \times \mathcal{T}$ es una relación binaria entre transacciones e itemsets.

Definición 1.11. (Conexión de Galois) Sea $D = (\Gamma, \mathcal{T}, \mathcal{R})$ un contexto de minería de datos. Para $T \subseteq \Gamma$ e $I \in \mathcal{T}$ se definen las dos aplicaciones $f : 2^{\Gamma} \longrightarrow 2^{\mathcal{T}}$, $g : 2^{\mathcal{T}} \longrightarrow 2^{\Gamma}$ donde

$$f(T) = \{i \in \mathcal{T} : \forall t \in T, t\mathcal{R}i\}$$
$$g(I) = \{t \in \Gamma : \forall i \in I, t\mathcal{R}i\}$$

El par (f, g) forma una *conexión de Galois* entre los conjuntos 2^{Γ} y $2^{\mathcal{T}}$ que son las partes[3] de los conjuntos Γ y $\mathcal{T}$ respectivamente.

[3]Las partes de un conjunto X notado en este caso por 2^X ($\mathcal{P}(X)$ en otros casos) es el conjunto de todos los posibles subconjuntos de X.

En general una conexión de Galois (f, g) cumple las siguientes propiedades:

- Si $I_1 \subseteq I_2$ entonces $g(I_2) \subseteq g(I_1)$.
- Si $T_1 \subseteq T_2$ entonces $f(T_2) \subseteq f(T_1)$.
- Si $T \subseteq g(I)$ entonces $I \subseteq f(T)$.

Definición 1.12. (Operador de clausura de Galois) Dada una conexión de Galois (f, g), se definen los *operadores de clausura* $h = f \circ g : 2^{\mathcal{T}} \longrightarrow 2^{\mathcal{T}}$ y $h' = g \circ f : 2^{\Gamma} \longrightarrow 2^{\Gamma}$. Dados $I, I_1, I_2 \subseteq \mathcal{T}$ y $T, T_1, T_2 \subseteq \Gamma$ los operadores h y h' cumplen las propiedades enumeradas a continuación:

1. Extensión: $I \subseteq h(I)$, $T \subseteq h'(T)$.
2. Idempotencia: $h(h(I)) = h(I)$, $h'(h'(T)) = h'(T)$.
3. Monotonía: Si $I_1 \subseteq I_2$, entonces $h(I_1) \subseteq h(I_2)$. Si $T_1 \subseteq T_2$, entonces $h'(T_1) \subseteq h'(T_2)$.

Una vez definidos los operadores de clausura, la *clausura* de un itemsets I no es más que $h(I)$.

Definición 1.13. (Itemsets Cerrados) Un itemsets será *cerrado* si coincide con su clausura, en términos de los operadores definidos anteriormente,

$$I \text{ es cerrado } \Leftrightarrow h(I) = I.$$

Definición 1.14. (Itemsets Frecuentes y Cerrados) Un itemsets se dirá que es *frecuente y cerrado* si es cerrado y además su soporte es mayor o igual que el umbral predefinido *minsop*.

Proposición 1.1. *[Pasquier et al., 1998]*

1. *Todos los subconjuntos de un itemsets frecuente son también frecuentes.*
2. *Todos los conjuntos de items que contienen a un itemsets infrecuente son también infrecuentes.*
3. *El soporte de un itemsets I es igual al soporte de su clausura:*

$$\text{sop}(I) = \text{sop}(h(I)) \quad \forall I \subseteq \mathcal{T}.$$

Definición 1.15. (Retículo de itemsets frecuentes y cerrados) Sea $\mathcal{C}$ el conjunto de itemsets frecuentes y cerrados derivados de un contexto D usando el operador clausura h. Al par $(\mathcal{C}, \leq)$ se le llama *retículo* de itemsets frecuentes y cerrados. $(\mathcal{C}, \leq)$ tiene las siguientes propiedades:

1. La clausura de $(\mathcal{C}, \leq)$ es más simple que la estructura de itemsets frecuentes.
2. Todos los sub-itemsets que estén en la estructura también son frecuentes.
3. Desde este retículo se pueden obtener reglas de asociación.

Además la estructura dada por el retículo dota de un orden parcial a los elementos, de forma que dados $C_1, C_2 \in (\mathcal{C}, \leq)$,

$$C_1 \leq C_2 \Leftrightarrow C_1 \subseteq C_2.$$

Ejemplo 1.2. Consideremos el contexto de minería de datos D:

ID	Items
1	A C D
2	B C E
3	A B C E
4	B E
5	A B C E

Su retículo de itemsets frecuentes y cerrados lo podemos ver en la Figura 1.3. Observemos que los arcos unen (de abajo hacia arriba) a los itemsets mediante la operación de unión y a las transacciones mediante la de intersección. El elemento más bajo siempre es el vacío, y el elemento más alto es el conjunto de todos los items, aunque en este caso tiene soporte igual a cero.

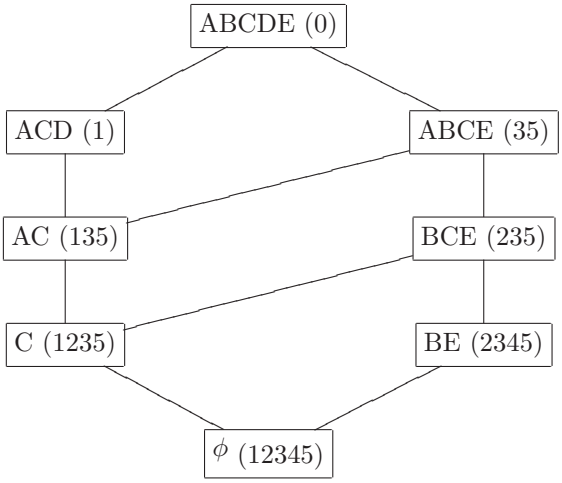

Figura 1.3: Retículo de itemsets frecuentes y cerrados.

El retículo de itemsets frecuentes y cerrados es una estructura que recoge todos los datos necesarios para la extracción de reglas de asociación. Veamos cuáles son los principales pasos para obtener reglas de asociación desde el retículo de itemsets cerrados.

1. Encontrar todos los itemsets frecuentes y cerrados en la base de datos D.
2. Determinar las reglas de asociación r de la forma: $I_1 \rightarrow I_2 - I_1$ tales que $I_1 \subset I_2$ donde I_1, I_2 son itemsets frecuentes y cerrados.

   Si I_1 e I_2 tienen la misma clausura diremos que la regla es exacta y la confianza será igual a 1. Si I_1 e I_2 tienen diferente clausura se dirá que la regla es aproximada.

En [Jia et al., 2003] se propone una primera aproximación de este modelo de representación para la extracción de reglas de asociación difusas. Al nuevo retículo lo llaman *retículo de itemsets difusos frecuentes y cerrados*, y lo que hacen es suavizar la forma de comprobar si un itemsets está en la clausura de otro itemsets, de forma que dos itemsets difusos están en la misma *clausura difusa* si el soporte de ambos está proximo (la diferencia de ambos soportes es menor que un umbral predefinido). Pero en principio difiere de la idea inicial de itemsets cerrado, puesto que la definición inicial compara si las transacciones en las que ambos itemsets ocurren son las mismas, mientras que en la versión difusa comparan el número de transacciones, es decir, el soporte.

Representación y Formulación mediante Tablas de Contingencia

La representación del conocimiento mediante una tabla de contingencia se utiliza en muchos ámbitos para representar la aparición conjunta de sucesos, valores de variables, etc. En nuestro campo también ha sido empleado con éxito ya que también aquí se trata de representar y evaluar las cuatro posibles combinaciones de dos itemsets y las correspondientes frecuencias [Gaines, 1991], [Tsumoto and Tanaka, 1995], [Zembowicz and Zytkow, 1996], [Ho and Scott, 1996], [Silverstein et al., 1998], [Yao and Zhong, 1999]. Este último trabajo presenta un análisis de las medidas más usadas para reglas de asociación del tipo $A \rightarrow B$ usando unas cantidades básicas que definen: la generalidad de A, el soporte absoluto de B dado A, el cambio de soporte de B dado A, el soporte mutuo de B y A; y el grado de independencia de A y B medido de dos formas distintas. Estas medidas no son más que las conocidas medidas para el soporte de A, la confianza de $A \rightarrow B$, $Conf(A \rightarrow B) - \text{sop}(A)$, $P(A \cap B)/P(A \cup B)$, y las dos medidas

de interés: $\frac{P(A\cap B)}{P(A)P(B)}$, y la dada por Piatetsky-Shapiro $P(A,B)-P(A)P(B)$ respectivamente. En la Tabla 1.3 quedan reflejadas las medidas y sus correspondencias con las medidas conocidas o bien en términos de probabilidades.

Nombre	**Correspondencia**
Generalidad de A	Soporte de A
Soporte absoluto de B	Confianza de $A \to B$
Cambio de soporte de B dado A	$\text{Conf}(A \to B) - \text{sop}(A)$
Soporte mutuo de B y A	$P(A \cap B)/P(A \cup B)$
Grado de independencia de A y B (I)	Interés: $P(A \cap B)/P(A)P(B)$
Grado de independencia de A y B (II)	Interés P-S: $P(A \cap B) - P(A)P(B)$

Tabla 1.3: Cantidades básicas tomadas en [Yao and Zhong, 1999].

Dichas medidas son clasificadas en tres grupos: generales (la medida involucra sólo una parte de la regla, es decir, el antecedente o bien el consecuente), asociación en un sentido (la medida no es simétrica) y asociación de doble sentido (la medida es simétrica).

En [Kodratoff, 2001] el autor se sirve también de una tabla de contingencia para explicar el funcionamiento y significado de distintos tipos de medidas. Para ello escoge diecinueve medidas distintas que analiza según su semántica. Representando las frecuencias mediante la siguiente tabla de contingencia

	B	$\neg B$
A	a	b
$\neg A$	c	d

podemos representar las medidas analizadas en [Kodratoff, 2001] en la siguiente tabla:

Nombre	**Definición**
Soporte Descriptivo	$\frac{a}{a+b+c+d}$
Soporte Causal	$\frac{a+d}{a+b+c+d}$
Confirmación Descriptiva	$\frac{a-b}{a+b+c+d}$
Confirmación Causal	$\frac{a+d-2b}{a+b+c+d}$
Confianza Descriptiva	$\frac{a}{a+b}$
Confirmación de la Confianza Descriptiva	$\frac{a-b}{a+b}$
Confianza Causal	$\frac{a}{2(a+b)} + \frac{d}{2(b+d)}$
Confirmación de la Confianza Causal	$\frac{a}{2(a+b)} + \frac{d}{2(b+d)} - \frac{b}{a+b}$
Convicción	$\frac{(a+b)(b+d)}{b(a+b+c+d)}$
Interés	$\frac{a(a+b+c+d)}{(a+b)(a+c)}$
Coeficiente de Correlación	$\frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$
Dependencia Descriptiva	$\left\lvert\frac{ad-bc}{(a+b)(a+b+c+d)}\right\rvert$ si $a+b \neq 0$
Dependencia Causal Putativa	$\frac{(a+2b+d)(ad-bc)}{(a+b)(b+d)(a+b+c+d)}$

Además de las medidas enumeradas en esta tabla, en [Kodratoff, 2001] se incluyen las medidas de "Dependencia Causal Falsa", "Confirmación Falsa", "Dependencia Causal Indirecta" y "Confirmación Indirecta" que hacen uso de un tercer item para calcular si la regla encontrada es falsa o bien es indirecta. De esta forma si tenemos un retículo de items podremos calcular cuáles son realmente las asociaciones verdaderas y eliminar aquellas que sean falsas o que dependan de un tercer factor (indirectas).

Formulación mediante Ejemplos Positivos y Negativos

La idea de usar una tabla de contingencia para generalizar conceptos crisp a difusos cuando manejamos reglas de asociación ha sido usada en varias ocasiones. En [Dubois et al., 2003], [Sudkamp, 2005], [Cock et al., 2005] la tabla de contingencia viene dada mediante una clasificación de las transacciones en ejemplos positivos, negativos e irrelevantes. Los trabajos [Dubois et al., 2003], [Sudkamp, 2005] usan esta terminología, mientras que en [Cock et al., 2005] realizan una división con intersecciones no vacías en ejemplos positivos, no-positivos, negativos y no-negativos. Explicaremos de forma más detallada en qué consisten ambas propuestas.

Dubois et al. clasifican las transacciones de una base de datos en tres tipos a partir de una regla de asociación $A \to B$ fijada de antemano [Dubois et al., 2003] obteniendo una partición de la base de datos:

$$\begin{aligned} S_+ &:= \{(x,y) \,|\, x \in A \wedge y \in B\} \\ S_- &:= \{(x,y) \,|\, x \in A \wedge y \notin B\} \\ S_\pm &:= \{(x,y) \,|\, x \notin A\} \end{aligned} \tag{1.13}$$

donde S_+, S_- y $S_\pm$ denotan los ejemplos *positivos, negativos* e *irrelevantes* respectivamente. Las medidas de soporte y confianza son expresadas en términos de los cardinales de dichos conjuntos. Después esta clasificación se extiende al caso difuso, de forma que si A y B son conjuntos difusos, entonces S_+, S_- y $S_\pm$ son también difusos, notando por $S_+(x,y)$ el grado de pertenencia del punto (x,y) en el conjunto S_+ de los ejemplos positivos, y de forma similar para S_- y $S_\pm$. Esto crea varias maneras de extenderlos al caso difuso, por ejemplo para definir $S_\pm$ se puede utilizar

$$(x,y) \in S_\pm \Leftrightarrow \neg(x \in A),$$

mientras que para S_- y S_+ existen varias opciones. Una de ellas sería:

$$\begin{aligned} (x,y) \in S_+ &:= (x \in A) \wedge (y \in B), \\ (x,y) \in S_- &:= (x \in A) \wedge \neg(y \in B). \end{aligned} \tag{1.14}$$

Y otra alternativa sería:

$$\begin{aligned} (x,y) \in S_+ &:= (x \in A) \wedge (y \in B), \\ (x,y) \in S_- &:= \neg((x \in A) \Rightarrow (y \in B)). \end{aligned} \tag{1.15}$$

donde $\Rightarrow$ es una implicación difusa (ver definición A.3). Observemos que desde el punto de vista lógico formal, las expresiones (1.14) y (1.15) son equivalentes, pero en el caso difuso dependerán de los operadores que elijamos para la negación y para la implicación difusa.

En el caso crisp, se cumple que

$$|S_+| + |S_-| + |S_\pm| = |D|\,, \tag{1.16}$$

pero para garantizar que se cumpla también cuando A y B son difusos, en [Dubois et al., 2003] demuestran que si la negación difusa es la definida como $\alpha \longmapsto 1 - \alpha$; el problema se reduce a encontrar una t-norma, $\otimes$, que cumpla la siguiente igualdad

$$S_+(x,y) + S_-(x,y) + S_\pm(x,y) = 1 \qquad \forall\,(x,y)$$

siendo

$$S_+(x,y) := A(x) \otimes B(y)$$
$$S_-(x,y) := A(x) \otimes (1 - B(y))$$
$$S_\pm(x,y) := 1 - A(x).$$

Como conclusión en [Dubois et al., 2003] se obtiene que la t-norma debe ser el producto y la implicación $\alpha \Rightarrow \beta$ es de la forma $(1-\alpha) + \alpha \otimes \beta$.

En [Sudkamp, 2005] se usa también la clasificación dada en el caso anterior y se hace un estudio de varias medidas para la validación de reglas de asociación mediante dicha partición de la base de datos D. Los autores también se centran en el análisis de las anomalías surgidas por el uso de bordes cuando se usa una partición crisp del dominio de los atributos para reglas cuantitativas; introduciendo el enfoque dado en [Dubois et al., 2003] para hacer una partición difusa de D mediante los conjuntos difusos S_+, S_- y $S_\pm$, y concluyendo también que el producto es la única t-norma que permite una partición con mejores propiedades.

El tercer enfoque ofrecido en [Cock et al., 2005] difiere bastante de los anteriores puesto que no ofrece una partición de la base de datos D, sino que hace una clasificación no disjunta de D dada la regla de asociación $A \to B$ como sigue:

$$\begin{aligned} \text{Ejemplos positivos } & \{(x,y) \,|\, x \in A \wedge y \in B\} \\ \text{Ejemplos no-positivos } & \{(x,y) \,|\, x \notin A \vee y \notin B\} \\ \text{Ejemplos negativos } & \{(x,y) \,|\, x \in A \wedge y \notin B\} \\ \text{Ejemplos no-negativos } & \{(x,y) \,|\, x \notin A \vee y \in B\} \end{aligned} \tag{1.17}$$

Y utiliza los cardinales de los conjuntos definidos en (1.17) para discernir la semántica de un grupo de medidas para el cumplimiento de reglas de asociación difusas.

1.2.2. Modelo Formal para Reglas de Asociación

El origen del modelo, que utilizaremos en el resto del documento, se remonta a una serie de trabajos sobre un método originado en Praga a mediados de los sesenta [Hájek et al., 1966], [Hájek and Havránek, 1978], y desarrollado hasta la actualidad en diversos campos. Su nombre es GUHA[4], y su principal objetivo es *"permitir al ordenador generar y evaluar todas las hipótesis que pueden ser*

[4]Siglas de General Unary Hypotheses Automaton

interesantes desde el punto de vista de los datos y del problema que queramos estudiar", [Hájek et al., 2004]. Una característica muy importante de GUHA es su fundamento lógico y estadístico que veremos en las próximas secciones y que nos ayudará a tener una mejor comprensión de la naturaleza de las reglas de asociación tal y como las manejamos ahora, así como de las medidas que se emplean en este ámbito, dando un amplio abanico de ellas con sus características y propiedades fundamentales.

Este modelo ha sido recientemente adaptado para el estudio de las reglas de asociación en varios trabajos [Rauch and Šimunek, 2000], [Rauch, 2005]. La notación que estableceremos está basada en la de esos trabajos pero difiere en el ambiente de trabajo (consideraremos bases de datos binarias) y en la definición de los *cuantificadores*, que introduciremos a continuación, cambiada para que se adapte mejor a la dinámica de las reglas de asociación.

El punto de partida es una base de datos binaria D donde las filas y las columnas representarán las transacciones y los items respectivamente (véase la Tabla 1.4). Como caso particular, los items podrán ser pares de la forma $\langle atributo, valor\rangle$ o $\langle atributo, intervalo\rangle$ dependiendo de cada base de datos particular, y su valor en la transacción t_i será 1 si el item se satisface en t_i o 0 en caso contrario.

D	i_1	i_2	$\ldots$	i_j	i_{j+1}	$\ldots$	i_n
t_1	1	0	$\ldots$	0	1	$\ldots$	0
t_2	0	1	$\ldots$	1	1	$\ldots$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
t_n	1	1	$\ldots$	0	1	$\ldots$	1

Tabla 1.4: Ejemplo de base de datos binaria D.

Para este modelo, un *itemset*, será una agregación de items atómicos mediante alguno o varios de los operadores lógicos $\wedge, \vee, \neg$. Un ejemplo de itemsets podría ser $i_1 \wedge (i_3 \vee \neg i_2)$, aunque el operador más utilizado en el ámbito de las reglas de asociación es la conjunción, veremos que la negación jugará un papel muy importante cuando extendamos el modelo para extraer reglas excepcionales y anómalas (Sección 2.4).

Con esta notación, una regla de asociación en este modelo [Rauch, 2005] será una expresión del tipo $\varphi \approx \psi$ donde φ y ψ son itemsets en el sentido

anterior y $\approx$ será un *cuantificador*. Dicho cuantificador estará asociado con φ y ψ por medio de una tabla de contingencia de cuatro huecos (4 fold table) que notaremos por $\mathcal{M}$ en la base de datos D. De forma gráfica la podemos representar por $\mathcal{M} = 4ft(\varphi, \psi, D) = \langle a, b, c, d\rangle$ o bien de la siguiente forma:

$\mathcal{M}$	ψ	$\neg\psi$
φ	a	b
$\neg\varphi$	c	d

siendo a el número de transacciones que satisfacen ψ y φ; b las que satisfacen φ y $\neg\psi$, etc. Además se debe verificar que $a + b + c + d > 0$. Y para quitar los casos triviales impondremos también que

$$a + b > 0, \qquad c + d > 0, \qquad a + c > 0, \qquad b + d > 0.$$

Observemos que $a + c$ es el número de veces que aparece ψ y $a + b$ el número de veces que aparece φ. Cuando el cuantificador esté asociado a dicha tabla lo llamaremos *cuantificador-4ft*.

El significado intuitivo de la expresión $\varphi \approx \psi$ es que los itemsets φ y ψ están asociados como indique el cuantificador en cuestión. La capacidad expresiva de este tipo de reglas es más fuerte que el de la clásica regla de asociación $X \to Y$ con X, Y conjuntos de items, ya que, según elijamos el cuantificador, la fórmula $\varphi \approx \psi$ puede modelizar diversos tipos de asociación.

Un ejemplo de regla de asociación podría ser $i_1 \wedge i_4 \approx i_2 \wedge i_5$ y el grado de validez vendrá dado por el valor que tenga el cuantificador-4ft, $\approx$, elegido que a su vez dependerá de $4ft(\varphi, \psi, D)$.

Los cuantificadores se usan normalmente en dos fases. La primera consiste en el cálculo del valor del cuantificador asociado a los dos itemsets φ y ψ. Este valor suele proceder de la adaptación de alguna medida de interés o bien de alguna medida predefinida (medidas estadísticas, medidas de recuperación de información) que valore algún tipo de relación entre los dos itemsets. Ahora comienza la segunda fase en la que para extraer reglas de asociación interesantes es necesario hacer una criba entre aquellas que no superen los umbrales necesarios de soporte y de la medida asociada al cuantificador. Por tanto, en esta fase, el valor del cuantificador es acotado por un umbral impuesto por el usuario. También se suele imponer un umbral al mínimo número de transacciones que satisfagan ambos itemsets, o lo que es igual, la imposición de que el soporte supere el umbral *minsop*.

Mediante la ejecución de estas dos fases se recogen dos tipos de información muy importantes en la extracción y validación de reglas de asociación. La primera

es el cálculo del valor del cuantificador que notaremos por $\approx(a,b,c,d)$ que es lo que conocemos como el valor de la medida de interés. La segunda es el cumplimiento de dos condiciones que nos transforman la regla de asociación $\varphi \approx \psi$ en un predicado lógico que puede ser verdadero o falso dependiendo de si se satisfacen o no las condiciones impuestas.

Una regla de asociación $\varphi \approx \psi$ será cierta en D (o en $\mathcal{M} = 4ft(\varphi,\psi,D)$) de forma simbólica $Val(\varphi \approx \psi) = true$ si y sólo si las condiciones asociadas al cuantificador-4ft $\approx$ se satisfacen para $\mathcal{M}$.

Distintos tipos de asociación entre los itemsets φ y ψ se pueden expresar definiendo cuantificadores-4ft adecuados.

Ejemplos de Cuantificadores-4ft

A continuación veremos algunos tipos de relaciones entre los itemsets φ y ψ que pueden ser expresados por cuantificadores-4ft. Se pueden encontrar estos ejemplos en [Havrànek, 1974], [Hájek and Havránek, 1978], [Hájek et al., 2004] y [Rauch, 2005].

- El cuantificador-4ft $\Rightarrow_I$ (*implicación*) se define como

$$\Rightarrow_I (a,b,c,d) = \frac{a}{a+b}. \tag{1.18}$$

  Observemos que $\Rightarrow_I$ es la conocida *confianza* [Agrawal et al., 1996] de la regla $\varphi \rightarrow \psi$. Este cuantificador nos dirá que $Val(\varphi \Rightarrow_I \psi) = true$ si y sólo si se cumplen las siguientes dos condiciones

$$\Rightarrow_I (a,b,c,d) \geq p \ \wedge \ a \geq Base$$

  donde $0 < p \leq 1$ es el umbral conocido como $minconf$ y $Base > 0$ es el mínimo número de transacciones satisfaciendo φ y ψ a la vez[5].

- El cuantificador-4ft $\sim$ (*asociación simple*) se define como

$$\sim (a,b,c,d) = ad - bc, \tag{1.19}$$

  y en este caso diremos que $Val(\varphi \sim \psi) = true$ si se cumple que

$$\sim (a,b,c,d) \geq 0 \ \wedge \ a \geq Base.$$

[5] Observemos que *Base* es un umbral para el soporte de la regla de asociación cuando éste viene dado en número en vez de en porcentaje.

Observemos que la primera condición es equivalente a probar que el determinante de la matriz asociada a la tabla $\mathcal{M}$ es positivo. En el caso de tener $\sim(a,b,c,d) = 0$ tendríamos la independencia entre los dos itemsets.

- El cuantificador-4ft $\Rightarrow^!$ (*implicación crítica*) es

$$\sum_{i=a}^{a+b} \binom{a+b}{i} p^i (1-p)^{a+b-i}$$

donde $0 < p \leq 1$. Diremos que $Val(\varphi \Rightarrow^! \psi) = true$ si se cumple

$$\Rightarrow^! (a,b,c,d) \leq \alpha \ \wedge \ a \geq Base$$

con $0 < \alpha <$0.5 y $Base > 0$. La segunda condición es la misma que en los casos anteriores, pero la primera condición corresponde al test estadístico al nivel α de la hipótesis nula $H_0 : P(\psi|\varphi) \leq p$ y con hipótesis alternativa $H_1 : P(\psi|\varphi) > p$ donde $P(\psi|\varphi)$ es la probabilidad condicionada de ψ bajo la condición de ocurrencia de φ.

- El cuantificador-4ft $\Leftrightarrow$ (*implicación doble*) es

$$\Leftrightarrow (a,b,c,d) = \frac{a}{a+b+c}.$$

De modo análogo a los casos anteriores, diremos que $Val(\varphi \Leftrightarrow \psi) = true$ si

$$\Leftrightarrow (a,b,c,d) \geq p \ \wedge \ a \geq Base$$

donde $0 < p \leq 1$ y $Base > 0$. Esta condición significa que al menos hay una proporción p de transacciones satisfaciendo φ ó ψ que satisfacen a la vez φ y ψ, y que hay al menos $Base$ transacciones de D satisfaciendo φ y ψ.

- El cuantificador-4ft $\equiv$ (*equivalencia*) es

$$\equiv (a,b,c,d) = \frac{a+d}{a+b+c+d}.$$

Diremos que $Val(\varphi \equiv \psi) = true$ si

$$\equiv (a,b,c,d) \geq p \ \wedge \ a \geq Base.$$

donde $0 < p \leq 1$ y $Base > 0$. Esto significa que φ y ψ tienen el mismo valor (ambas verdaderas o ambas falsas) en una proporción de al menos p entre todas las transacciones de D y que hay al menos $Base$ transacciones satisfaciendo φ y ψ.

- El cuantificador-4ft $\equiv^!$ (*equivalencia crítica*)

$$\equiv^! (a,b,c,d) = \sum_{i=a+d}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

donde $0 < p \leq 1$ y $n = a+b+c+d$. Diremos que $Val(\varphi \equiv^! \psi) = true$ si

$$\equiv^! (a,b,c,d) \leq \alpha \ \wedge \ a \geq Base$$

con $0 < \alpha < 0.5$ y $Base > 0$.

- El cuantificador-4ft $\sim_{Eqc}$ (*equicardinalidad*) viene dado por

$$\sim_{Eqc} (a,b,c,d) = (a+b) - (a+c).$$

Diremos que $Val(\varphi \sim_{Eqc} \psi) = true$ si

$$\sim_{Eqc} (a,b,c,d) = 0.$$

Este cuantificador indica que la regla es válida si los cardinales de φ y ψ son el mismo. Esto implica que también coincidan los cardinales de $\neg\varphi$ y $\neg\psi$. En particular que $4ft(\varphi, \psi, D)$ es simétrica (esto es, $b = c$).

- El cuantificador-4ft $\sim_F$ (*cuantificador de Fisher*) es

$$\sim_F (a,b,c,d) = \sum_{i=a}^{\text{mín}(r,k)} \frac{\binom{k}{i}\binom{n-k}{r-i}}{\binom{r}{n}}$$

donde hemos llamado $r = a+b$ y $k = a+c$. Diremos que $Val(\varphi \sim_F \psi) = true$ si

$$\sim_F (a,b,c,d) \leq \alpha \ \wedge \ ad > bc \ \wedge \ a \geq Base,$$

con $0 < \alpha \leq 0.5$ y $Base > 0$. Estas condiciones corresponden al test estadístico al nivel α para la hipótesis nula de independencia de φ y ψ contra la hipótesis alternativa de dependencia.

Existen muchos cuantificadores-4ft [Rauch, 2005], [Hájek and Havránek, 1978] con distintos tipos de propiedades. Según la semántica que recojan y la forma en la que estén definidos se pueden clasificar atendiendo a varios criterios. Veamos a continuación algunas de las clases más interesantes.

Clases de cuantificadores-4ft

Según el tipo de cuantificador-4ft que usemos para extraer la regla de asociación, conseguiremos distintos tipos de reglas de asociación. De esta forma la clasificación de los cuantificadores induce una clasificación de las reglas de asociación [Rauch, 2005]. Veamos las clases más importantes de cuantificadores-4ft [Hájek and Havránek, 1978], [Iványek, 1999].

Definición 1.16. Sean $a, b, c, d, a', b', c', d'$ frecuencias de dos tablas-4ft arbitrarias $\langle a', b', c', d' \rangle$ y $\langle a, b, c, d \rangle$, entonces

(1) un cuantificador $\approx (a, b)$ es **implicacional,** $\approx \in I$, si siempre que

$$a' \geq a \quad \wedge \quad b' \leq b$$

se cumple que $\approx (a', b') \geq \approx (a, b)$,

(2) un cuantificador $\approx (a, b, c)$ es **doble implicacional,** $\approx \in DI$, si siempre que

$$a' \geq a \quad \wedge \quad b' \leq b \quad \wedge \quad c' \leq c$$

se cumple que $\approx (a', b', c') \geq \approx (a, b, c)$,

(3) un cuantificador $\approx (a, b, c)$ es **Σ-doble implicacional,** $\approx \in \Sigma DI$, si siempre que

$$a' \geq a \quad \wedge \quad b' + c' \leq b + c$$

se cumple que $\approx (a', b', c') \geq \approx (a, b, c)$,

(4) un cuantificador $\approx (a, b, c, d)$ es **equivalente,** $\approx \in E$, si siempre que

$$a' \geq a \quad \wedge \quad b' \leq b \quad \wedge \quad c' \leq c \quad \wedge \quad d' \geq d$$

se cumple que $\approx (a', b', c', d') \geq \approx (a, b, c, d)$,

(5) un cuantificador $\sim (a, b, c, d)$ es **Σ-equivalente,** $\approx \in \Sigma E$, si siempre que

$$a' + d' \geq a + d \quad \wedge \quad b' + c' \leq b + c$$

se cumple que $\approx (a', b', c', d') \geq \approx (a, b, c, d)$.

Cuando el cuantificador-4ft no dependa de todos los valores (a, b, c, d) si no sólo de algunos, quitaremos aquellos que sean innecesarios para simplificar la notación como hemos hecho en la anterior definición.

Las condiciones impuestas en la definición nos dice que la tabla-4ft asociada $\mathcal{M}' = \langle a', b', c', d' \rangle$ a los dos itemsets es en algún sentido "mejor" desde el punto de vista implicacional, equivalente, etc. que la tabla-4ft $\mathcal{M} = \langle a, b, c, d \rangle$. Por ejemplo la condición para los cuantificadores implicacionales ($a' \geq a \wedge b' \leq b$) nos dice que $\mathcal{M}'$ es mejor desde el punto de vista implicacional que $\mathcal{M}$ puesto que D' tiene más transacciones satisfaciendo φ y ψ que D; y D' tiene menos transacciones satisfaciendo φ y $\neg\psi$ que D.

En [Hájek and Havránek, 1978] y [Rauch, 2005] se puede ver la prueba formal de la inclusión de los siguientes cuantificadores-4ft en las clases definidas anteriormente:

- $\sim$ es equivalente pero no implicacional.
- $\Rightarrow_I \in I$.
- $\Rightarrow^! \in I$.
- $\Leftrightarrow \in DI \cap \Sigma DI$.
- $\Leftrightarrow^! \in \Sigma DI \cap DI$.
- $\equiv \in E$.
- $\equiv^! \in \Sigma E \cap E$.

Por otro lado, estas clases de cuantificadores-4ft están relacionadas mediante la siguiente proposición.

Proposición 1.2. *[Iványek, 1999] Se cumplen las siguientes inclusiones:*

- $I \subset DI \subset E$,
- $\Sigma DI \subset DI$, *y*
- $\Sigma E \subset E$.

Además dichas inclusiones son estrictas, ya que por ejemplo, $\sim \in E$ pero $\sim \notin I$.

Hay otras clases de cuantificadores-4ft con interesantes propiedades teóricas y prácticas [Hájek and Havránek, 1978], [Rauch, 2005]. Un ejemplo es la clase de los cuantificadores *simétricos*, cuya motivación es la propiedad de simetría de algunos tipos de reglas de asociación. Una regla con esta propiedad nos dirá que $\varphi \approx \psi$ es cierta si y sólo si la regla $\psi \approx \varphi$ es cierta, en términos de $\mathcal{M}$ equivaldría a decir que si la tabla $4ft(\varphi, \psi, D) = \langle a, b, c, d \rangle$ entonces la tabla $4ft(\psi, \varphi, \mathcal{D})$ es $\langle a, c, b, d \rangle$.

Estas consideraciones nos llevan a la siguiente definición:

Definición 1.17. Un cuantificador-4ft $\approx$ es *simétrico* si

$$\approx(a,b,c,d) = \approx(a,c,b,d).$$

Ejemplos de cuantificadores simétricos son $\Leftrightarrow$, $\Leftrightarrow^!$, $\equiv$, $\equiv^!$, el cuantificador de Fisher $\sim_F$ y el cuantificador $\sim$.

Otra clase interesante es la clase de los cuantificadores-4ft con la F-propiedad. La clase de los cuantificadores que cumplen la F-propiedad engloba a una clase de cuantificadores que tienen las mismas propiedades teóricas que el cuantificador de Fisher $\sim_F$.

Definición 1.18. Un cuantificador-4ft $\approx$ tiene la *F-propiedad* si satisface las dos propiedades siguientes

1. Si $\approx(a,b,c,d) = 1$ y $b \geq c-1 \geq 0$ entonces $\approx(a,b+1,c-1,d) = 1$.
2. Si $\approx(a,b,c,d) = 1$ y $c \geq b-1 \geq 0$ entonces $\approx(a,b-1,c+1,d) = 1$.

Definición 1.19. Sea $\approx$ un cuantificador equivalente. Diremos que es *saturable* si cumple las siguientes propiedades:

1. Para cualquier matriz $\langle a,b,c,d\rangle$ con $d \neq 0$, existe $a' \geq a$ tal que $\approx(a',b,c,d) = 1$.
2. Para cualquier matriz $\langle a,b,c,d\rangle$ con $a \neq 0$, existe $d' \geq d$ tal que $\approx(a,b,c,d') = 1$.
3. Para cualquier matriz $\mathcal{M} = \langle a,b,c,d\rangle$, hay otra matriz $\mathcal{M}'$ que contiene a $\mathcal{M}$ tal que $\approx(\mathcal{M}') = 0$.

Por ejemplo el cuantificador-4ft $\sim_F$ es saturable.

Otra clase de cuantificadores-4ft con un significado interesante que utilizaremos más adelante es el siguiente:

Definición 1.20. [Hájek et al., 2004] Un cuantificador-4ft $\approx$ se dirá que es *comparativo* si $Val(\varphi \approx \psi) = true$ implica que $ad > bc$, o equivalentemente si $a/(a+b) > (a+c)/(a+b+c+d)$.

Un cuantificador comparativo medirá el grado con el que el antecedente y el consecuente están positivamente asociados y su valor será cero si el antecedente y el consecuente son estadísticamente independientes. Un ejemplo de cuantificador comparativo es $\sim$.

La siguiente sección presenta un modelo para representar y operar con propiedades difusas distinto a la Teoría de Subconjuntos Difusos propuesto por Zadeh. Este modelo propone el uso de niveles de restricción y extiende las operaciones usuales entre conjuntos de forma que las propiedades lógicas (crisp) usuales sigan cumpliéndose, cosa que no ocurría en el modelo de Zadeh. Además, este modelo será útil en el capítulo siguiente para proporcionar un modelo formal, que extiende al presentado aquí, para reglas de asociación difusas.

1.3. Representación de Cantidades Imprecisas mediante Niveles de Restricción

Los conjuntos difusos son uno de los mejores modelos para representar y razonar con propiedades difusas (ver Apéndice A). Su desarrollo y aplicación en muy diversas áreas han dado muy buenos resultados. Esto ha sido posible porque las operaciones clásicas entre conjuntos han sido extendidas apropiadamente. Generalmente hay diferentes formas de extender las definiciones y las operaciones clásicas del caso crisp al difuso. Este es el caso de la intersección, la unión y el complemento que dependen de cómo se definan las t-normas, t-conormas y las negaciones difusas, siendo las posibilidades innumerables. Pero en el proceso de extensión al caso difuso, algunas de las propiedades clásicas entre conjuntos se pierden. Este es uno de los problemas de la lógica difusa: cómo definir las extensiones de la forma más adecuada posible, qué propiedades deberían cumplirse y qué semántica encierran estas nuevas definiciones. Este tipo de problema sale a la luz frecuentemente cuando las operaciones y definiciones que se extienden usan la negación, es decir, el complemento de un conjunto. Otro problema de la negación es la extensión de la implicación como $\neg A \vee B$ y el correspondiente conjunto de interpretaciones como grado de inclusión (en particular, las definiciones originales de inclusión e igualdad de un conjunto difuso son crisp).

Estos problemas, y otros, son muy importantes en algunas aplicaciones, por ejemplo cuando operamos con cardinales de conjuntos difusos. La implicación también es crucial para el razonamiento difuso y para aquellos conceptos que la utilizan como es el caso de la transitividad de relaciones difusas, i.e., $(R(a,b) \wedge R(b,c)) \Rightarrow R(a,c)$.

La propuesta de Sánchez et al. [Sánchez et al., 2008b] para representar propiedades imprecisas mediante niveles de restricción extiende las operaciones usuales del caso crisp al difuso cumpliendo las propiedades lógicas usuales entre dichas operaciones. Esta propuesta se basa en la conocida idea de que una propiedad difusa en un universo X puede ser representada por una colección de realizaciones crisp. Como caso particular, cuando el conjunto es difuso, dichas realizaciones crisp son los conocidos α-cortes, y serán los llamados *niveles de restricción* (por brevedad se usará la notación RL). Este modelo es más general que el de los conjuntos difusos, aunque estos jugarán un papel importante puesto que ayudarán a obtener la representación por niveles de restricción de las propiedades atómicas.

Este modelo servirá como herramienta en el siguiente capítulo para establecer un modelo formal para reglas de asociación difusas que extienda al presentado en la sección 1.2. Usando dicho modelo desarrollaremos un método para extender las medidas de interés para reglas difusas.

1.3.1. Definiciones

Definición 1.21. [Sánchez et al., 2008b] Un *RL-set* Λ es un conjunto finito de niveles de restricción $\Lambda = \{\alpha_1, \ldots, \alpha_m\}$ verificando $1 = \alpha_1 > \alpha_2 > \cdots > \alpha_m > \alpha_{m+1} = 0$ para $m \geq 1$.

En general, el *RL*-set de una propiedad atómica representada mediante un conjunto difuso A está definido como sigue:

Definición 1.22. [Sánchez et al., 2008b] Sea A un conjunto difuso definido sobre el referencial X. Entonces el *RL*-set para A viene dado por:

$$\Lambda_A = \{A(x) \,|\, x \in X\} \cup \{1\} \tag{1.20}$$

donde $A(x)$ es el grado de pertenencia de x al conjunto difuso A.

El *RL*-set empleado para representar una propiedad difusa se obtiene como la unión de los *RL*-sets de las propiedades atómicas que definan dicha propiedad.

Para representar una propiedad difusa en X mediante niveles de restricción usaremos una *RL-representación* definida como un par (Λ, ρ) donde Λ es un *RL*-set y

$$\rho : \Lambda \to \mathcal{P}(X) \tag{1.21}$$

es una función que aplica cada nivel de restricción en su realización crisp, que representará la propiedad imprecisa en dicho nivel. Como ejemplo, la *RL*-representación de una propiedad atómica imprecisa definida por un conjunto difuso A será el par (Λ_A, ρ_A), donde Λ_A viene dada por la ecuación (1.20) y $\rho_A(\alpha) = A_\alpha = \{x \in X \,|\, A(x) \geq \alpha\}$ para todo $\alpha \in \Lambda_A$.

Dada una propiedad difusa P representada por (Λ_P, ρ_P), se define el conjunto de las *representaciones crisp* de P como

$$\Omega_P = \{\rho_P(\alpha) | \alpha \in \Lambda_P\}. \tag{1.22}$$

Para una propiedad atómica A, el conjunto de representaciones crisp Ω_A es justamente el conjunto de los α-cortes de A. Sin embargo, observemos que en la definición de *RL*-representación no hay ninguna restricción sobre las posibles

representaciones crisp de las propiedades que no sean atómicas. En particular, como consecuencia de las operaciones, las RL-representaciones no tienen que estar anidadas, por tanto, las RL-representaciones finales no coinciden siempre con la representación mediante α-cortes de los conjuntos difusos. Para definir nuevas propiedades no atómicas mediante operaciones es conveniente extender la función ρ para cualquier $\alpha \in (0, 1]$.

Definición 1.23. [Sánchez et al., 2008b] Sea (Λ, ρ) una RL-representación con $\Lambda = \{\alpha_1, \ldots, \alpha_m\}$. Sea $\alpha \in (0, 1]$ y $\alpha_i, \alpha_{i+1} \in \Lambda$ cumpliendo que $\alpha_i > \alpha > \alpha_{i+1}$. Entonces definiremos

$$\rho(\alpha) = \rho(\alpha_i). \tag{1.23}$$

Si nos fijamos en esta definición, esta extensión para valores que no estén en el RL-set de la función ρ, es la natural si pensamos en un conjunto difuso A y en sus α-cortes. Usando esta definición se definirá la equivalencia entre dos RL-representaciones.

Definición 1.24. [Sánchez et al., 2008b] Sean (Λ, ρ) y (Λ', ρ') dos RL-representaciones sobre X. Diremos que ambas representaciones son *equivalentes* y lo notaremos por $(\Lambda, \rho) \equiv (\Lambda', \rho')$, si y sólo si, $\forall \alpha \in (0, 1]$

$$\rho(\alpha) = \rho(\alpha'). \tag{1.24}$$

1.3.2. Operaciones Lógicas

En esta sección veremos las operaciones lógicas de disyunción, conjunción y negación, que nos harán falta en el próximo capítulo para el desarrollo de un modelo lógico para las reglas de asociación difusas. Las ideas básicas de cómo se definen dichas operaciones se encuentran en [Sánchez et al., 2008b].

Definición 1.25. [Sánchez et al., 2008b] Sean P, Q dos propiedades difusas con RL-representaciones (Λ_P, ρ_P), (Λ_Q, ρ_Q). Entonces, $P \wedge Q$, $P \vee Q$ y $\neg P$ son propiedades difusas representadas por $(\Lambda_{P\wedge Q}, \rho_{P\wedge Q})$, $(\Lambda_{P\vee Q}, \rho_{P\vee Q})$ y $(\Lambda_{\neg P}, \rho_{\neg P})$ respectivamente, donde

$$\begin{aligned} \Lambda_{P\wedge Q} &= \Lambda_{P\vee Q} = \Lambda_P \cup \Lambda_Q \\ \Lambda_{\neg P} &= \Lambda_P \end{aligned} \tag{1.25}$$

y, para todo $\alpha \in (0, 1]$,

$$\begin{aligned} \rho_{P\wedge Q}(\alpha) &= \rho_P(\alpha) \cap \rho_Q(\alpha), \\ \rho_{P\vee Q}(\alpha) &= \rho_P(\alpha) \cup \rho_Q(\alpha), \\ \rho_{\neg P}(\alpha) &= \overline{\rho_P(\alpha)}, \end{aligned} \tag{1.26}$$

donde $\overline{Y}$ es el complemento usual de un conjunto crisp Y.

Proposición 1.3. *[Sánchez et al., 2008b] Las operaciones $\wedge, \vee, \neg$ entre RL-representaciones verifican las propiedades ordinarias de equivalencia lógica como $\neg\neg A \equiv A$, las leyes de Morgan (una de ellas $\neg(A \wedge B) \equiv (\neg A \vee \neg B)$) y la ley del tercero excluido que puede ser expresada como $A \wedge \neg A \equiv \perp$ ó $A \vee \neg A \equiv \top$, donde $\top$ y $\perp$ son las propiedades atómicas que representan respectivamente la tautología (cuya RL-representación se obtiene del referencial X) y la contradicción (cuya representación se obtiene de ϕ).*

1.3.3. Números-RL

Basándose en las RL-representaciones y sus operaciones, Sánchez et al. definen los *números-RL* como una representación de cantidades difusas [Sánchez et al., 2008c]. Esta propuesta tiene dos grandes ventajas:

1. Los números-RL son representaciones de cantidades difusas que se obtienen fácilmente extendiendo las medidas habituales crisp a los conjuntos difusos y, en general, a RL-representaciones.

2. Las operaciones aritméticas y lógicas entre números-RL son extensiones únicas e inmediatas de las operaciones con números crisp, verificando las propiedades usuales del caso crisp. Además, la imprecisión no se incrementa necesariamente cuando realizamos operaciones, esta puede incluso disminuir.

Las siguientes definiciones y propiedades provienen de [Sánchez et al., 2008c]:

Definición 1.26. [Sánchez et al., 2008c] Un número RL-real es un par $(\Lambda, \mathcal{R})$ donde Λ es un RL-set y $\mathcal{R} : (0, 1] \to \mathbb{R}$.

Notaremos por $\mathbb{R}_{RL}$ al conjunto de los números RL-reales. El número RL-real R_x es una representación de un número real x si, y sólo si $\forall \alpha \in \Lambda_{R_x}$, $\mathcal{R}_{R_x}(\alpha) = x$. A dicho número RL-real lo notaremos por R_x o, de forma equivalente, x, ya que en el caso crisp el conjunto Λ_{R_x} no es importante. Las operaciones entre números RL-reales se extienden como sigue:

Definición 1.27. [Sánchez et al., 2008c] Sea $f : \mathbb{R}^n \to \mathbb{R}$ y sean $R_1, \ldots, R_n$ números RL-reales. Entonces $f(R_1, \ldots, R_n)$ es un número RL-real con

$$\Lambda_{f(R_1,\ldots,R_n)} = \bigcup_{1 \leq i \leq n} \Lambda_{R_i} \tag{1.27}$$

y, $\forall \alpha \in \Lambda_{f(R_1,\ldots,R_n)}$

$$\mathcal{R}_{f(R_1,\ldots,R_n)}(\alpha) = f(\mathcal{R}_{R_1}(\alpha),\ldots,\mathcal{R}_{R_n}(\alpha)) \tag{1.28}$$

Es evidente que las operaciones entre números RL-reales son extensiones consistentes de las operaciones usuales crisp. Además, las operaciones que no estén definidas para algunos valores reales, consecuentemente no estarán definidas para los números RL-reales que contengan dichos valores para alguno de sus niveles de restricción. Este es el caso de la división por 0, por lo que, R/R' estará definido sí y sólo si $0 \notin \Omega_{R'}$.

Para definir un orden en los números-RL, en [Sánchez et al., 2008a] se define el concepto de orden-RL.

Definición 1.28. Un *orden-RL* es un par (Λ, δ) donde Λ es un RL-set y

$$\delta : (0,1] \longrightarrow \{=, <, >\}.$$

Observemos que un orden-RL es un orden difuso basado en los órdenes posibles entre las representaciones asociadas a los números-RL. El orden-RL nos permite extender los procedimientos basados en ordenar cuando usemos niveles de restricción [Sánchez et al., 2008a].

Definición 1.29. [Sánchez et al., 2008a] El orden de dos números RL-reales R_1 y R_2 es un orden-RL $(\Lambda_{R_1,R_2}, \delta_{R_1,R_2})$ donde $\Lambda_{R_1,R_2} = \Lambda_{R_1} \cup \Lambda_{R_2}$ y para todo $\alpha \in \Lambda_{R_1,R_2}$,

$$\delta_{R_1,R_2}(\alpha) = \begin{cases} = & \text{si } \mathcal{R}_{R_1}(\alpha) = \mathcal{R}_{R_2}(\alpha) \\ < & \text{si } \mathcal{R}_{R_1}(\alpha) < \mathcal{R}_{R_2}(\alpha) \\ > & \text{si } \mathcal{R}_{R_1}(\alpha) > \mathcal{R}_{R_2}(\alpha). \end{cases} \tag{1.29}$$

Diremos que $R_1 = R_2$ (respectivamente $R_1 < R_2, R_1 > R_2$) si, y sólo si, $\forall \alpha \in \Lambda_{R_1,R_2}$, se cumple que $\mathcal{R}_{R_1}(\alpha) = \mathcal{R}_{R_2}(\alpha)$ (respectivamente $\mathcal{R}_{R_1}(\alpha) < \mathcal{R}_{R_2}(\alpha), \mathcal{R}_{R_1}(\alpha) > \mathcal{R}_{R_2}(\alpha)$).

1.3.4. Probabilidad basada en números-RL

Notaremos por $[0,1]_{RL} \subseteq \mathbb{R}_{RL}$ al conjunto de todos los números RL-reales cumpliendo que $0 \leq R \leq 1$, $\forall R \in [0,1]_{RL}$. Definiremos los conceptos de RL-espacio de probabilidad y RL-probabilidad como en [Sánchez et al., 2008a]:

Definición 1.30. [Sánchez et al., 2008a] Un RL-espacio de probabilidad es una terna (X, Σ, P) donde

1. X es un conjunto crisp.

2. Σ es una colección de eventos difusos definidos por sus RL-representaciones sobre X, cerrado para las operaciones de complemento y para las uniones numerables de eventos. Σ siempre contendrá al propio universo X.

3. $P : \Sigma \to [0,1]_{RL}$ verifica los Axiomas de Kolmogorov, i.e., $P(X) = 1$, $P(E) \geq 0$, $\forall E \in \Sigma$, y para cualquier colección finita de representaciones disjuntas $E_1, \ldots, E_n$,

$$P(E_1 \cup \cdots \cup E_n) = \sum_{i=1}^{n} P(E_i). \tag{1.30}$$

En la anterior definición, P es una medida de RL-probabilidad. En particular, las medidas de RL-probabilidad pueden obtenerse fácilmente de las medidas ordinarias de probabilidad. Sea (U, F, p) un espacio de probabilidad crisp y sea U_F^{RL} el conjunto de todas las RL-representaciones en U tales que para cualquier $A \in U_F^{RL}$, $\Omega_A \subseteq F$. Usando las propiedades de F, obtenemos que U_F^{RL} es cerrado para el complemento y para las uniones numerables. Además cumple que $U \in U_F^{RL}$. Para cualquier $A \in U_F^{RL}$, sea $P(A) \in [0,1]_{RL}$ definido por

$$\rho_{P(A)}(\alpha) = p\left(\rho_A(\alpha)\right) \qquad \forall \alpha \in \Lambda_A. \tag{1.31}$$

Se puede comprobar fácilmente que (U, U_F^{RL}, P) es un RL-espacio de probabilidad y P es una medida de RL-probabilidad.

También es sencillo comprobar que las RL-probabilidades verifican las siguientes propiedades:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ P(\neg A) &= 1 - P(A). \end{aligned} \tag{1.32}$$

Además, usando la definición 1.30, es evidente que las probabilidades (crisp) son un caso particular de las RL-probabilidades ya que los eventos crisp son casos particulares de los eventos imprecisos descritos por sus RL-representaciones.

Es sencillo demostrar que se preservan las relaciones aritméticas usuales entre probabilidades en cada nivel de restricción, y, por tanto, se preservan las propiedades usuales entre RL-probabilidades. Por ejemplo, $P(A) + P(\neg A) = 1$, es decir, $\forall \alpha \in \Lambda_A$,

$$p(\rho_A(\alpha)) + p(\rho_{\neg A}(\alpha)) = 1.$$

A continuación definimos el concepto de RL-probabilidad condicional como podemos encontrar en [Sánchez et al., 2008a]:

Definición 1.31. Sean A y B dos eventos imprecisos definidos en U por sus RL-representaciones $A = (\Lambda_A, \rho_A)$ y $B = (\Lambda_B, \rho_B)$ con $\rho_{P(B)}(\alpha) > 0$, $\forall \alpha \in (0, 1]$. La RL-probabilidad condicionada de A dado B en U es

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \tag{1.33}$$

Proposición 1.4. *[Sánchez et al., 2008a] Sean A y B dos eventos imprecisos definidos en U por las RL-representaciones $A = (\Lambda_A, \rho_A)$ y $B = (\Lambda_B, \rho_B)$ con $\rho_{P(B)}(\alpha) > 0$, $\forall\, \alpha \in (0, 1]$. Entonces*

$$\rho_{P(A|B)}(\alpha) = \frac{\rho_{P(A \wedge B)}(\alpha)}{\rho_{P(B)}(\alpha)}. \tag{1.34}$$

Ejemplo 1.3. Sean X e Y dos propiedades difusas atómicas definidas en $U = \{u_1, \ldots, u_6\}$ por los siguientes conjuntos difusos:

$$\begin{aligned} X &= 1/u_1 + 0.8/u_2 + 0.5/u_3 + 0.4/u_5 \\ Y &= 0.9/u_1 + 0.6/u_3 + 0.5/u_4 \end{aligned} \tag{1.35}$$

Entonces, $\Lambda_X = \Lambda_{\neg X} = \{1, 0.8, 0.5, 0.4\}$ y $\Lambda_Y = \Lambda_{\neg Y} = \{1, 0.9, 0.6, 0.5\}$. El RL-set para cualquier operación entre X e Y será $\Lambda_X \cup \Lambda_Y$. La Tabla 1.5 muestra las RL-representaciones de X, $\neg X$, Y, $\neg Y$, $X \wedge Y$, $X \vee Y$ y $X \wedge \neg Y$. Si asumimos que cada u_i es equiprobable, en la Tabla 1.6 tenemos las RL-probabilidades correspondientes a los eventos imprecisos $P(X)$, $P(\neg X)$, $P(Y)$, $P(\neg Y)$, $P(X \wedge Y)$, $P(X \vee Y)$, $P(X \wedge \neg Y)$ y $P(Y|X)$. Observemos que la RL-probabilidad condicionada $P(X|Y)$ no está definida puesto que $\rho_{P(Y)}(1) = 0$.

α_i	$\rho_X(\alpha)$	$\rho_{\neg X}(\alpha)$	$\rho_Y(\alpha)$	$\rho_{\neg Y}(\alpha)$
1	$\{u_1\}$	$\{u_2, u_3, u_4, u_5, u_6\}$	ϕ	U
0.9	$\{u_1\}$	$\{u_2, u_3, u_4, u_5, u_6\}$	$\{u_1\}$	$\{u_2, u_3, u_4, u_5, u_6\}$
0.8	$\{u_1, u_2\}$	$\{u_3, u_4, u_5, u_6\}$	$\{u_1\}$	$\{u_2, u_3, u_4, u_5, u_6\}$
0.6	$\{u_1, u_2\}$	$\{u_3, u_4, u_5, u_6\}$	$\{u_1, u_3\}$	$\{u_2, u_4, u_5, u_6\}$
0.5	$\{u_1, u_2, u_3\}$	$\{u_4, u_5, u_6\}$	$\{u_1, u_3, u_4\}$	$\{u_2, u_5, u_6\}$
0.4	$\{u_1, u_2, u_3, u_5\}$	$\{u_4, u_6\}$	$\{u_1, u_3, u_4\}$	$\{u_2, u_5, u_6\}$

α_i	$\rho_{X \wedge Y}(\alpha)$	$\rho_{X \vee Y}(\alpha)$	$\rho_{X \wedge \neg Y}(\alpha)$
1	ϕ	$\{u_1\}$	$\{u_1\}$
0.9	$\{u_1\}$	$\{u_1\}$	ϕ
0.8	$\{u_1\}$	$\{u_1, u_2\}$	$\{u_2\}$
0.6	$\{u_1\}$	$\{u_1, u_2, u_3\}$	$\{u_2\}$
0.5	$\{u_1, u_3\}$	$\{u_1, u_2, u_3, u_4\}$	$\{u_2\}$
0.4	$\{u_1, u_3\}$	$\{u_1, u_2, u_3, u_4, u_5\}$	$\{u_2, u_5\}$

Tabla 1.5: *RL*-representaciones asociadas a algunas operaciones entre las propiedades imprecisas dadas por X e Y.

α_i	$\rho_{P(X)}(\alpha)$	$\rho_{P(\neg X)}(\alpha)$	$\rho_{P(Y)}(\alpha)$	$\rho_{P(\neg Y)}(\alpha)$
1	1/6	5/6	0	1
0.9	1/6	5/6	1/6	5/6
0.8	1/3	2/3	1/6	5/6
0.6	1/3	2/3	1/3	2/3
0.5	1/2	1/2	1/2	1/2
0.4	2/3	1/3	1/2	1/2

α_i	$\rho_{P(X\wedge Y)}(\alpha)$	$\rho_{P(X\vee Y)}(\alpha)$	$\rho_{P(X\wedge\neg Y)}(\alpha)$	$\rho_{P(Y\|X)}(\alpha)$
1	0	1/6	1/6	0
0.9	1/6	1/6	0	1
0.8	1/6	1/3	1/6	1/2
0.6	1/6	1/2	1/6	1/2
0.5	1/3	2/3	1/6	2/3
0.4	1/3	5/6	1/3	1/2

Tabla 1.6: *RL*-probabilidades en U de las propiedades imprecisas de la Tabla 1.5 y la *RL*-probabilidad condicional $P(Y|X)$.

1.4. Resumen

A lo largo de este capítulo, hemos recordado los importantes conceptos generales de extracción del conocimiento (KD), minería de datos (DM) y regla de asociación (AR) que sientan las bases sobre las que trabajará el resto de la memoria. En particular, el concepto clave es el de regla de asociación con sus múltiples extensiones y aplicaciones que serán de utilidad en el resto de la memoria.

También se han presentado dos secciones claves para el desarrollo de esta memoria. La primera de ellas es la búsqueda de un modelo formal para la representación y evaluación de las reglas de asociación, corazón de este documento. El elegido debido a sus buenas propiedades tanto para modelar la estructura como el comportamiento de las reglas de asociación ha sido el basado en GUHA.

La segunda sección, clave en esta memoria, es la introducción de un modelo para la representación de la imprecisión mediante niveles de restricción que resuelve algunos problemas derivados del uso de la negación que ofrecía la teoría de subconjuntos difusos (ver Apéndice A).

Capítulo 2

Resultados Teóricos

Este capítulo contiene los resultados desarrollados para cubrir el principal objetivo de esta tesis: el desarrollo de un modelo formal para la representación y evaluación de reglas de asociación. Siguiendo con el desarrollo del modelo lógico, en primer lugar pondremos de manifiesto la relación existente entre dicho modelo y las propiedades que debe cumplir una 'buena' medida de interés. Después extenderemos el modelo para reglas difusas y propondremos un procedimiento para extender las medidas de interés crisp para reglas difusas, y por último, usaremos el modelo para estudiar y obtener otro tipo de reglas, como son las reglas excepcionales o reglas anómalas. Además proveeremos algunos resultados desarrollados de forma colateral que complementan el trabajo realizado.

El capítulo está organizado como se detalla a continuación:

La sección 2.1 repasa las distintas propiedades o principios que debería cumplir una buena medida de interés establecidos hasta el momento introduciendo dos nuevos principios que consideramos deberían satisfacerse también. Después establecemos una relación entre dichos principios y los tipos de cuantificadores-4ft definidos en la sección 1.2 y finalizamos introduciendo el factor de certeza [Berzal et al., 2001] como un cuantificador-4ft demostrando que es equivalente, y que por tanto, cumple los principios definidos para una buena medida de interés.

La sección 2.2 contiene la extensión del modelo GUHA para reglas de asociación difusas utilizando como herramienta el modelo de representación mediante niveles de restricción visto en la sección 1.3. Vía dicha extensión proponemos un método para extender de forma natural las medidas de interés (en nuestro caso,

cuantificadores) para la extracción de reglas difusas.

La sección 2.3 presenta diversos métodos para extraer conocimiento en bases de datos formadas por bolsas usando tanto reglas de asociación difusas como dependencias graduales. También ofrecemos una aplicación para obtener reglas de asociación en documentos de texto mediante su representación como bolsas difusas.

Para finalizar el capítulo, y por tanto, los resultados teóricos desarrollados en esta tesis, la sección 2.4 estudia la extracción de nuevos tipos de conocimiento excepcional y anómalo utilizando como herramienta las reglas de asociación. Repasamos las propuestas realizadas por otros autores y las comparamos a nuestras propuestas desde un punto de vista teórico utilizando para ello el modelo lógico. Además, definimos un nuevo tipo de reglas llamadas *reglas dobles* que junto con las reglas excepcionales y las anómalas ayudarán a analizar de forma más precisa la relación existente entre dos itemsets en una base de datos.

Queremos destacar que los resultados teóricos provistos en este capítulo se complementarán con los experimentos que se presentan en el capítulo siguiente.

2.1. Medidas de Interés y su Relación con el Modelo Formal

En el proceso de extracción de las reglas de asociación se han desarrollado varias líneas de trabajo para obtener distintas posibilidades de obtención de información con esta útil herramienta. Una de las líneas más prolíferas ha sido la que se encarga de buscar nuevos tipos de reglas como las que hemos comentado en la sección 1.1. Otra importante línea de investigación se centra en obtener distintas medidas de interés que ayuden en el proceso de extracción de las reglas. Entre estos tipos de medidas hay algunas que permiten filtrar o distinguir entre reglas útiles para el usuario, o aquellas que puedan resultar más interesantes desde algún punto de vista fijado "a priori". En este ámbito también se han desarrollado diversas propuestas para establecer un conjunto de *principios* o *propiedades* que debería cumplir toda "buena" medida de interés.

Esta sección pretende dar una visión global sobre el proceso que se sigue para valorar la "importancia" o el "interés" de las reglas de asociación extraídas en una base de datos, sobre las distintas medidas de interés propuestas hasta el momento y los principios que miden la bondad de dichas medidas.

Además estableceremos una conexión entre estos principios y el modelo formal que estamos desarrollando, proporcionando un procedimiento sencillo para comprobar cuándo una medida de interés, en nuestro caso cuándo un cuantificador-4ft, cumple los principios de bondad establecidos. También daremos como ejemplo una medida (el factor de certeza [Berzal et al., 2002]), cuyas buenas propiedades para extraer reglas de asociación se han probado en la práctica, para probar teóricamente si dicha medida cumple los principios de bondad establecidos.

2.1.1. Medidas de Interés para Reglas de Asociación

En Minería de Datos uno de los procesos más importantes es el de valorar adecuadamente el conocimiento obtenido. En particular, las medidas de interés nos ayudan a medir y cuantificar la utilidad, interés y valor de las reglas de asociación obtenidas en los datos.

Entre los procedimientos usados, podemos distinguir básicamente cuatro técnicas para obtener las que llamaremos *reglas interesantes* para el usuario. En la Figura 2.1 podemos ver de forma esquematizada dichas técnicas. En la primera, las reglas obtenidas en el proceso de minería son mostradas directamente al usuario. En (b) las reglas interesantes son obtenidas en un post-proceso de la minería (DM).

(c) muestra la búsqueda integrada de reglas de asociación en el proceso de DM. En la práctica el procedimiento que más se utiliza es el (c), pero de un tiempo a esta parte, se han desarrollado procedimientos como el esquema expuesto en (d) en los que la participación del usuario es una parte importante del proceso. Este es el caso del procedimiento presentado en [Liu et al., 1999a] donde se usa un post-análisis llamado Interestingness Analysis System (IAS) para ayudar al usuario a identificar reglas interesantes

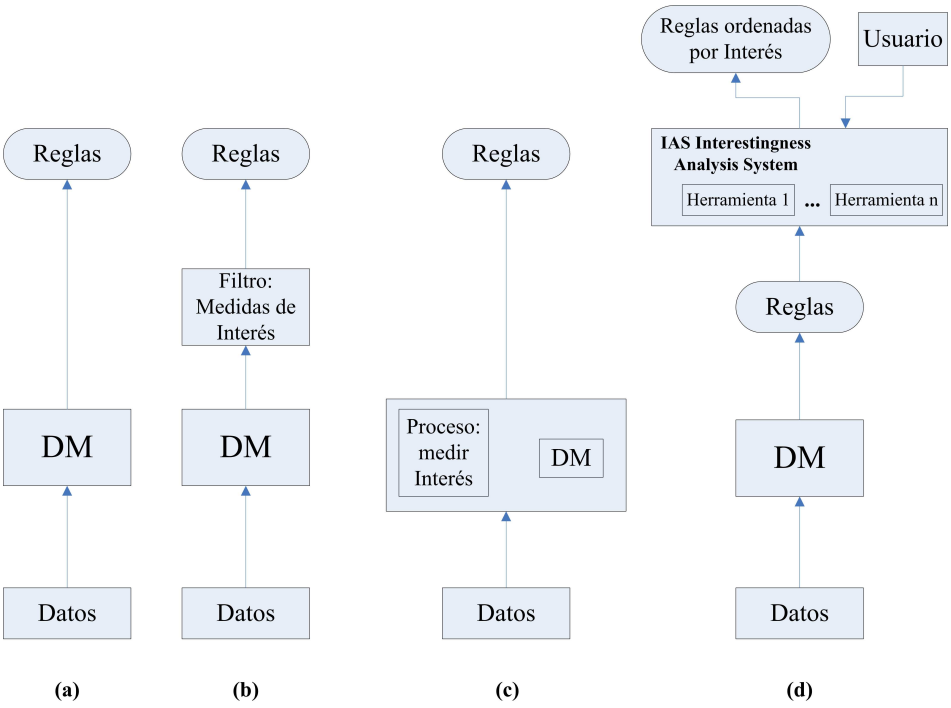

Figura 2.1: Técnicas para obtener reglas interesantes.

En [Tan and Kumar, 2000], [Hilderman and Hamilton, 2003], [McGarry, 2005], [Ohsaki et al., 2004], [Hébert and Crémilleux, 2006] y [Lallich et al., 2006] podemos encontrar varias recopilaciones sobre los distintos enfoques para medir el interés de una regla de asociación. En particular, [McGarry, 2005] ofrece una adecuada clasificación de los tipos de medidas de interés en tres categorías: *medidas de similitud*, muy utilizadas en clasificación, que sirven para comparar la semejanza entre objetos, transacciones e incluso reglas; las *medidas objetivas* basadas en medidas estadísticas o en propiedades de los patrones; y *medidas subjetivas* que

son aquellas derivadas de las creencias o expectativas del usuario respecto a su dominio particular del problema (Ver Figura 2.2).

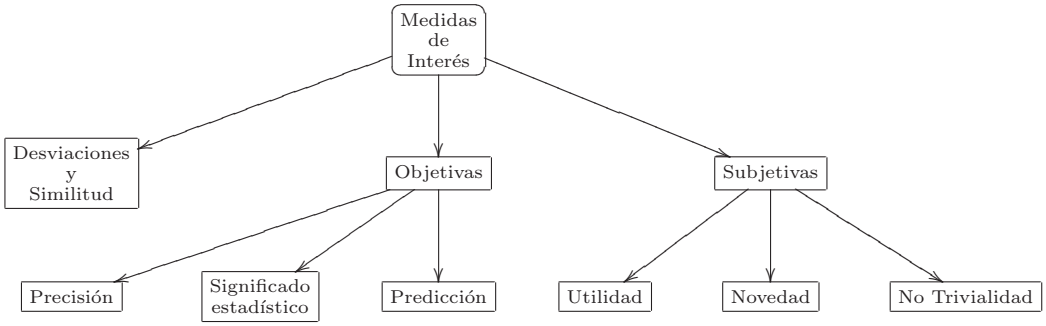

Figura 2.2: Criterios para medir el interés de las reglas de asociación

En [Frawley et al., 1992] se introduce el concepto de interés desde una perspectiva muy prometedora. Como ya hemos comentado, la extracción de conocimiento pretende hallar conocimiento útil y comprensible para el usuario. Para ello queremos utilizar las reglas de asociación, una herramienta muy extendida en el ámbito de la minería de datos. En particular pretendemos extraer reglas de asociación interesantes. Una regla será interesante cuando sea novedosa, útil y no trivial de calcular. La novedad de una regla dependerá de la referencia que tengamos, por ejemplo, las creencias del usuario. Para que sea útil deberá ayudar a conseguir algún objetivo del usuario; y la no trivialidad, se relaciona con una cierta autonomía del proceso y de la evaluación de los resultados; aunque también podríamos identificarlo con el concepto de redundancia. De forma poco rigurosa, una regla es trivial cuando se puede obtener utilizando otras reglas ya obtenidas con anterioridad. La Figura 2.2[1] muestra un gráfico con los distintos tipos de medidas de interés y algunos de los aspectos tenidos en cuenta para definir dichas medidas.

Observemos que todas las propiedades señaladas para las medidas subjetivas son necesarias para medir si el conocimiento es interesante o no, ya que por ejemplo aunque una regla sea inesperada o novedosa, esto no quiere decir que sea válida. Incluso si una regla es inesperada y válida puede pasar que no sea útil. Todas estas propiedades pueden cambiar para distintos usuarios ya que son conceptos altamente subjetivos como se indicaba en el esquema anterior. Pero no es tarea fácil implementar una métrica o criterio para descubrir los patrones más interesantes ya

[1] El gráfico está ligeramente modificado del original en [McGarry, 2005].

que éstos son los imprevistos y sorprendentes, es decir, un patrón útil o novedoso se sabe que lo es cuando lo ves, pero es difícil proponer un camino para descubrirlo [Gaines, 1996].

Respecto a los otros dos grupos de medidas, las de similitud (también llamadas medidas de desviación o diferencia) ofrecen métricas consolidadas que suelen utilizarse sobre todo en técnicas de agrupamiento y clasificación. Y las objetivas se concentran en dar un grado de confianza de las reglas descubiertas midiendo diversos factores que no requieren la participación del usuario. La mayoría de estas medidas provienen de métodos estadísticos que ayudan a medir tanto la representatividad de la regla, su capacidad de predicción, la correlación entre los items involucrados, etc.

2.1.2. Criterios para Medir el Interés

Los criterios mostrados en la Figura 2.2 son algunos de los más utilizados para medir el interés de una regla de asociación. Algunos de ellos han dado lugar al desarrollo de herramientas específicas para encontrar de manera óptima un conjunto de reglas cumpliendo la propiedad elegida. En particular, analizaremos tres enfoques basados en la utilidad, la novedad y la redundancia respectivamente.

Utilidad para medir el interés

Es difícil predecir la utilidad de una regla y definir una medida para ello. Silberschatz y Tuzhilin (1996) propusieron dividir el espacio de las reglas encontradas en clases de equivalencia, asociando a cada clase un tipo de acción [Silberschatz and Tuzhilin, 1996]. En [Shen et al., 2002] y [Chan et al., 2003] se puede encontrar una técnica para obtener reglas de acuerdo a unos objetivos prefijados por el usuario. Utilizan un modelo al que llaman OOA (Objective-Oriented utility-based Association) para obtener las reglas 'útiles' utilizando una adaptación del algoritmo Apriori. En [Bade et al., 2006] se define también una función de utilidad a partir de los conocimientos expresados por el usuario. Maximizando dicha función proponen un método alternativo para alcanzar mejores resultados en el caso particular de la clasificación jerárquica.

Ahora bien, puesto que la utilidad es un concepto altamente subjetivo, éste puede venir dado o influenciado por varios factores que normalmente son útiles para el usuario, y por tanto, se podría enfocar la búsqueda de una *medida de utilidad* teniendo en cuenta la acción conjunta de, por ejemplo, diversos factores: si son necesarias algunas restricciones para que se cumpla la regla, el periodo de

tiempo de utilidad de la regla, si la regla tiene efectos laterales, si la información obtenida puede utilizarse en el momento, etc.

La novedad y el interés

Algunos autores definen novedad como '*una hipótesis H es novedosa, con respecto a un conjunto de creencias B, si y sólo si H no es derivable de B*'. Siguiendo esta definición, una regla que contradiga el conjunto de creencias será sorprendente y novedosa.

Si encontramos una regla inesperada que contradiga alguna creencia del usuario, deberemos sorprendernos. Padmanabhan y Tuzhilin han trabajado sobre el problema de encontrar reglas inesperadas y el conflicto que supone con el conocimiento y las creencias del usuario [Padmanabhan and Tuzhilin, 1998], [Padmanabhan and Tuzhilin, 1999], [Padmanabhan and Tuzhilin, 2002] . En trabajos posteriores [Padmanabhan, 2004] y [Padmanabhan and Tuzhilin, 2006] utilizan también los conceptos de "novedad" y "minimalidad" de forma conjunta para obtener el conjunto mínimo de reglas inesperadas para el usuario. Para ello se basan en la *contradicción lógica* para generar las reglas inesperadas para el usuario y en la *redundancia* para generar conjuntos minimales de reglas.

Otros autores [Colton et al., 2005] también se centran en la novedad, la verosimilitud y la no-trivialidad como criterios comunes de su sistema de extracción.

Hay que tener cuidado cuando extraemos reglas inesperadas. Entre los problemas que nos podemos encontrar, es la ocurrencia de paradojas, problema que ha sido investigado por diversos autores. En particular, la paradoja de Simpson ha sido identificada por muchos investigadores en el ámbito de la minería de datos.

Por ejemplo, Fabris y Freitas [Fabris and Freitas, 1999] desarrollaron un algoritmo para detectar esta paradoja y después calcular su magnitud, usando el grado de sorpresa. En [Padmanabhan, 2004] el concepto de minimalidad débil es utilizado para desarrollar un método práctico para eliminar algunas paradojas que surgen usando ciertas medidas de interés.

Otros autores como Zhong y Yao [Zhong and Yao, 2003] hacen una interpretación un poco distinta del concepto de inesperado, centrándolo en la peculiaridad. Este enfoque conlleva la búsqueda de reglas con bajos valores de soporte que satisfagan unos umbrales mínimos para una función que mide la peculiaridad de los datos. Zhong propone además una aproximación usando lógica difusa para la expresión y la interpretación de estas reglas peculiares.

La detección de novedad en distintos registros de datos es muy importante en técnicas de agrupamiento para obtener una buena clasificación. [Markou and Singh, 2003] contiene una visión general de los trabajos más importantes que utilizan técnicas estadísticas para encontrar novedad en los datos.

Trivialidad y Redundancia

Cuando se extraen reglas de asociación el usuario espera obtener reglas no triviales. Normalmente en el proceso de extracción se exige que ambos lados de la regla sean disjuntos para evitar que surjan reglas que contengan en el antecedente y en el consecuente los mismos items, lo que daría lugar a reglas triviales. Algunos trabajos tratan la trivialidad desde el punto de vista de la redundancia de un conjunto de reglas. Este conjunto se considerará minimal y no redundante si desde cualquier conjunto que lo contenga se pueden obtener el conjunto total de reglas de interés.

Otros autores definen el concepto de redundancia de una regla con respecto a otra, mientras que muy pocos trabajan con una definición de redundancia basada en un conjunto de reglas [Padmanabhan and Tuzhilin, 2006]. Otros trabajos definen la redundancia teniendo en cuenta la estructura de las reglas o bien usando una medida de interés. El principal problema de tomar la estructura de la regla para definir el proceso de inferencia es que generalmente no se puede calcular el valor de la medida de interés de las reglas inferidas. Por eso, muchos autores han optado por hacer una aproximación conjunta teniendo en cuenta tanto la estructura de la regla como la medida de interés en cuestión.

Como ejemplos de procesos de inferencia para obtener el conjunto minimal de reglas no redundantes podemos nombrar algunos basados en el concepto de *closed itemsets* y *frequent closed itemsets* (ver sección 1.2.1), como en los trabajos [Zaki, 2000], [Zaki, 2004] de Zaki, en los que se usa la transitividad y el aumento para derivar el conjunto de reglas de asociación desde un pequeño conjunto de reglas. Un enfoque parecido se sigue en el trabajo de Bastide et al. en [Bastide et al., 2000]. Jaroszewicz y Simovici presentan un método para desechar reglas redundantes usando un principio de máxima entropía [Jaroszewicz and Simovici, 2002].

2.1.3. Tipos de Medidas de Interés

En esta sección analizaremos con más detalle los tres tipos de medidas de interés presentados en la Figura 2.2, describiendo las propuestas más representativas en

cada uno de ellos.

Desviaciones y Medidas de Similitud

Las llamadas *medidas de similitud* intentan medir y cuantificar la diferencia o la igualdad entre la información capturada por dos objetos. En particular, se puede destacar el uso de estas medidas en el ámbito de las reglas de clasificación [Roubos et al., 2000] y también en Recuperación de Información [Martinovic et al., 2008]. Otros trabajos a destacar que usan desviaciones y medidas de similitud son [Piatetsky-Shapiro and Matheus, 1994], [Roddick and Rice, 2001], [Roussinov and Zhao, 2003].

Medidas de Interés Objetivas

Las medidas de interés *objetivas* son aquellas que no tienen en consideración la participación del usuario, es decir, únicamente utilizan los datos almacenados para cuantificar el interés o la importancia de la regla extraída. Como se mostraba en la Figura 2.2 principalmente hay tres características que se han tenido en cuenta para definir distintas medidas de interés objetivas: la precisión, el significado estadístico y la predicción.

Muchas de las medidas objetivas se definen para medir el grado de precisión que posee la regla de asociación, como por ejemplo la frecuencia de aparición de los items en las transacciones. También se han considerado el significado estadístico de la regla asociándolo con la medida de interés. Por ejemplo, el test estadístico de independencia χ^2 se ha usado para medir la dependencia entre dos itemsets.

Otros de los enfoques seguidos es el de ver cuál es la capacidad de predicción de la regla, es decir, si nos sirve para conseguir nuevas reglas, o bien, si con ella se pueden predecir reglas que ocurran en un futuro.

Además de estos aspectos, hay también otros muchos que se han utilizado para medir el interés de una regla de asociación de forma objetiva, como son: la fuerza de la regla, su cobertura (también relacionado con la predicción), su precisión, la confianza que podemos depositar en ella, etc. Consultar [Hilderman and Hamilton, 2003] y [Tan and Kumar, 2000], para un análisis más detallado. También se han propuesto cuáles deberían ser las características deseables que toda medida de interés objetiva debería tener. Esto lo analizaremos detalladamente en la sección 2.1.4.

A continuación veremos algunas de las medidas objetivas propuestas hasta el momento por orden cronológico comentando sus características más notables.

Major y Mangano usan un *factor de predominio* [Major and Mangano, 1995] que ordena las reglas según el lugar que ocupan en una jerarquía, puntuando las reglas según sean potencialmente interesantes, técnicamente interesantes y genuinamente interesantes. Para ello, tienen en cuenta varios criterios como el de ejecución, simplicidad y significado. Las reglas genuinamente interesantes (GI) por ejemplo, son aquellas que son simples, más generales y no redundantes. El proceso que utilizan es dar tres medidas, una por cada tipo de regla, y después el experto usa estas medidas para eliminar las reglas que no crea necesarias o que sean poco interesantes.

También se han hecho otros trabajos con reglas de asociación que contienen datos temporales. Ramaswamy et al. crearon una medidas para identificar reglas de asociación en series temporales de datos [Ramaswamy et al., 1998]. Esta propuesta es útil para detectar tendencias o patrones que ocurren en periodos concretos de tiempo. También desarrollaron la medida objetiva llamada *lift* para determinar la importancia de cada regla de asociación.

Wang et al. [Wang et al., 1998] propusieron usar la *medida-J* de Smyth y Goodman [Smyth and Goodman, 1992] y otra basada en la confianza para medir el interés de las reglas.

Hussain et al. [Hussain et al., 2000] utilizan medidas para la entropía para identificar reglas excepcionales.

La medida de interés propuesta en [Korn et al., 2000] evalúa la calidad de las reglas usando un 'guessing error'(GE) referido a la estimación de los valores que faltan en la base de datos.

Otros autores hacen uso de métricas fundamentales, como por ejemplo, Ikizler & Gvenir [Ikizler and Gvenir, 2001]. Gago & Bento [Gago and Bento, 1998] derivan las medidas de interés de métricas para medir la distancia entre dos reglas para seleccionar aquellas que sirvan para predecir otras reglas; ya que normalmente las reglas más interesantes son aquellas que no se pueden predecir del conocimiento existente.

Para reducir el número de reglas extraídas siendo estas más interesantes, en [Berzal et al., 2002] se propuso usar el Factor de Certeza que analizaremos con más profundidad en la sección 2.1.7.

Tan et al. realizaron una revisión de 20 medidas de interés analizándolas según la estructura interna de los datos, su coste computacional y los efectos de escala observados [Tan et al., 2002] .

Jaroszewicz y Simovici utilizan redes Bayesianas para evaluar el interés de conjuntos de items frecuentes [Jaroszewicz and Simovici, 2004]. En este caso el

criterio de interés viene dado por el valor absoluto de la diferencia entre el soporte de la regla y el soporte calculado por la red Bayesiana.

En la tabla 2.1 podemos ver varios ejemplos de medidas de interés descritas usando el modelo formal visto en 1.2, por lo que a representa el número de transacciones de la base de datos que contienen al antecedente y al consecuente a la vez, b las que contienen al antecedente pero no al consecuente, c las que contienen el consecuente y no al antecedente, y por último, d el número de transacciones que no contienen ni al antecedente ni al consecuente. Podemos encontrar un resumen de dichas medidas en [Lallich et al., 2006] y [Tan et al., 2004].

Medida	**Fórmula**	**Acrónimo**
Soporte	$\frac{a}{a+b+c+d}$	SUP
Confianza	$\frac{a}{a+b}$	CONF
Confianza Centrada	$\frac{ad-bc}{(a+b+c+d)(a+b)}$	CENCONF
Ganascia	$\frac{a-b}{a+b}$	GAN
Piatetsky-Shapiro	$\frac{ad-bc}{a+b+c+d}$	PS
Loevinger	$\frac{ad-bc}{(b+d)(a+b)}$	LOE
Zhang	$\frac{ad-bc}{\max\{a(b+d),b(a+c)\}}$	ZHANG
Coeficiente de Correlación	$\frac{ad-bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$	R
Índice de Implicación	$\frac{bc-ad}{\sqrt{(a+b+c+d)(a+b)(b+d)}}$	IMPIND
Lift o Interés	$\frac{a(a+b+c+d)}{(a+b)(a+c)}$	LIFT
Coseno	$\frac{a(a+b+c+d)}{\sqrt{(a+b)(a+c)}}$	IS
Mínima Contradicción	$\frac{a-b}{a+c}$	LC
Convicción	$\frac{(a+b)(b+d)}{b(a+b+c+d)}$	CONV
Sebag-Schoenauer	$\frac{a}{b}$	SEB
Factor de Bayes	$\frac{a(b+d)}{b(a+c)}$	BF
Odds Ratio	$\frac{ad}{bc}$	α

Tabla 2.1: Medidas de interés objetivas para reglas de asociación

Freitas [Freitas, 1998], [Freitas, 1999] reconoció el amplio abanico de medidas que había y propuso para las futuras medidas de interés unos ciertos criterios que

deberían tenerse en cuenta a la hora de definir una buena medida de interés:

- *El número de atributos.* Las reglas con pocos antecedentes son generalmente más fáciles de comprender y las reglas con muchos antecedentes pueden aparecer a veces debido al ruido de los datos o bien pueden ser un caso especial.
- *Coste de la mala clasificación.* Dependiendo del contexto en el que se trabaje, un falso positivo puede ser menos dañino que un falso negativo.
- *Clases de la distribución.* Los resultados dependerán de la clasificación o división de los datos hecha.
- *Orden de los atributos.* Algunos atributos serán mejores a la hora de discriminar entre clases, así si se encuentran dichos atributos en una regla, reflejará que ésta es más importante que el resto.
- *Asimetría de la medida.* Será mejor una medida que sepa distinguir entre antecedente y consecuente.

Además de las medidas objetivas, también están las medidas subjetivas que permiten la participación del usuario o incluso se pueden combinar ambas para obtener un conjunto de reglas interesantes según los criterios del usuario [Sinoara and Rezende, 2006]. En la siguiente sección veremos las propuestas más importantes usando medidas subjetivas.

Medidas de Interés Subjetivas

Las medidas objetivas no son suficientes para discernir entre las reglas más interesantes para cada usuario en el proceso de minería de datos. Sin embargo, las medidas subjetivas intentan solventar dicho inconveniente usando generalmente las creencias del usuario para ordenar los patrones descubiertos en el algoritmo de DM escogido en cada caso. Esto es lo que los autores llaman medir la novedad o lo inesperado de una regla. Además también se tienen en cuenta si la regla es útil y no trivial.

La novedad dependerá del sistema de referencia propuesto, que podemos convenir que sea el conjunto de creencias del usuario. La utilidad de la regla se medirá de acuerdo a un conjunto de objetivos que el usuario quiere conseguir. Y la no trivialidad la podríamos asociar a la parte no subjetiva del interés, es decir, que haya estadísticamente datos que sustenten la regla o bien que dicha regla no

pueda obtenerse de otras reglas obtenidas con anterioridad. Esto último podríamos conseguirlo si la regla es no redundante.

Liu et al. hacen un pequeño análisis sobre las medidas subjetivas de interés y consideran que éstas deben tener en cuenta si la regla es inesperada por el usuario (unexpectedness) y también si es útil (usefulness) para él [Liu et al., 1999a], [Liu et al., 2000]. Pero como ya hemos comentado, en realidad éstas serían las propiedades subjetivas que debería tener la regla, pero no hay que olvidar que también es importante la parte objetiva.

Silberschatz y Tuzhilin [Silberschatz and Tuzhilin, 1996] basan la identificación de reglas interesantes en si éstas son imprevistas y útiles, midiendo el interés de una regla en términos de cómo afecta al sistema de creencias del usuario, dando éste un grado de confianza a cada creencia.

Hay varias técnicas para concebir los sistemas de creencias del usuario y esto normalmente implica un ejercicio de adquisición del conocimiento del dominio de los expertos. En los trabajos hechos hasta ahora se han usado principalmente las siguientes herramientas para reconocer las reglas más interesantes usando el conocimiento subjetivo del usuario:

- Medidas probabilísticas. Se han utilizado aproximaciones Bayesianas para poder usar probabilidades condicionadas [Silberschatz and Tuzhilin, 1995], [Silberschatz and Tuzhilin, 1996].
- Medidas para la distancia sintáctica. Este enfoque se basa en la distancia entre las nuevas reglas y el conjunto de creencias. Por ejemplo, si los consecuentes de una regla son los mismos que los esperados por el usuario, pero los antecedentes son muy distintos, entonces la regla se consideraría interesante [Liu et al., 1999a].
- Contradicción lógica. Una medida estadística objetiva es usada para medir si la regla es esperada o no (mediante soporte y confianza por ejemplo), y después se analiza si hay alguna diferencia con los grados esperados por el usuario de dichas medidas [Padmanabhan and Tuzhilin, 1999], [Padmanabhan and Tuzhilin, 2002].

A continuación veremos un resumen de los trabajos más representativos que usan la participación del usuario para encontrar las reglas más interesantes.

Trabajo 2.1. User-expectation Method

Liu et al. proponen un método llamado 'user-expectation' donde se le pregunta al usuario qué reglas espera encontrar de acuerdo a su conocimiento o a sus sensaciones [Liu et al., 1999a]. Una vez almacenadas estas reglas, el sistema usa una técnica difusa para obtener aquellas reglas en contra de las expectativas del usuario y después las ordena de acuerdo a los resultados obtenidos.

Formalmente, sea D la base de datos. Una técnica de DM, T, se aplica a D para obtener las reglas de asociación. Un subconjunto B de estas reglas debe satisfacer los umbrales de soporte, confianza o de otras medidas estadísticas impuestas por el usuario. Sea I el conjunto de reglas interesantes (subjetivas) para el usuario, luego se tiene que $I \subseteq B$, y debemos tener en cuenta que:

- Dada una misma base de datos D y el conjunto de reglas B, distintos usuarios estarán interesados en distintos subconjuntos de B.
- No todas las reglas de I son igualmente interesantes. Cada una tendrá un grado distinto de interés.
- I debe ser un conjunto dinámico en el sentido de que el usuario puede estar interesado o no en una regla dependiendo de factores tales como la fecha.

En general, B es mucho más grande que I. Esto implica que muchas de las reglas encontradas por T no son interesantes o útiles. Lo deseable es que un sistema sólo dé al usuario el conjunto de reglas interesantes y que además las ordene por grado de interés.

Las reglas que se manejan en el método propuesto por Liu et al. serán de la forma

$$P_1 \wedge P_2 \wedge \ldots \wedge P_n \longrightarrow C$$

donde "$\wedge$" denota la conjunción, y P_i son proposiciones de la forma $\langle atributo\ OP\ valor \rangle$ donde *atributo* es el nombre del atributo en la base de datos, *valor* es un posible valor para *atributo*, y $OP \in \{=, \neq, <, >, \leq, \geq\}$ es el operador. C es el consecuente que tendrá la forma $\langle clase = valor \rangle$.

Recordemos que la aproximación utilizada en este caso es la que podemos observar en (d) de la Figura 2.1. En ella se usa un post-análisis llamado IAS (Interestingness Analysis System) para ayudar al usuario a identificar reglas interesantes.

En este procedimiento, primero deberemos saber las creencias del usuario, y de acuerdo a ellas evaluaremos el grado de conformidad con las reglas. Hay principalmente dos pasos en la técnica propuesta por estos autores:

Paso 1. El usuario propone las reglas que él espera encontrar en la base de datos D. Éstas podrán ser reglas difusas.

Paso 2. El sistema compara cada regla encontrada $B_i \in B$ con cada regla esperada por el usuario $E_i \in E$. Según el parecido de dichas reglas, se ordenan para dar al usuario las reglas más interesantes.

Para realizar el segundo paso, en [Liu et al., 1999a] se proponen cuatro algoritmos diferentes para ordenar las reglas y mostrárselas al usuario. Como ya hemos comentado, el usuario introduce su conocimiento por medio de reglas del mismo tipo que las obtenidas por el sistema, pero puede hacer uso de variables lingüísticas difusas.

Denotaremos por W_i el grado de compatibilidad entre la regla descubierta por el sistema $B_i \in B$ y el conjunto de reglas esperadas por el usuario E; y usaremos $w_{(i,j)}$ para medir dicho grado entre B_i y $E_j \in E$. Así dependiendo del valor de W_i ordenaremos las reglas en orden decreciente, siendo más interesantes aquellas con menor compatibilidad. Para calcular $w_{(i,j)}$ seguiremos dos pasos:

1. *Compatibilidad de los atributos.* En este paso se comparan los atributos que aparecen en B_i y en E_j. Para ello se almacenan en $A_{(i,j)}$ los atributos comunes a B_i y E_j y después se calcula

$$L_{(i,j)} = \frac{\left|A_{(i,j)}\right|}{\text{máx}(|e_j|\,|b_i|)}$$

donde $|e_j|$ y $|b_i|$ son el número de atributos en el antecedente de E_j y B_i respectivamente.

2. *Compatibilidad en el valor de los atributos.* Los autores denotan por $V_{(i,j)k}$ al grado de compatibilidad entre el valor del atributo a_k que se encuentra en B_i y E_j, y llaman $Z_{(i,j)}$ al grado de compatibilidad entre los valores del consecuente. Dependiendo de dichos valores se obtendrán las reglas más interesantes.

Como hemos mencionado, los valores $w_{(i,j)}$ pueden servir para ordenar las reglas encontradas conforme a las hipótesis y creencias del usuario. Así los autores proponen dos formas de calcular $w_{(i,j)}$ y por tanto de calcular W_i. Para más detalles consultar [Liu et al., 1999a].

Trabajo 2.2. IAS: Interestingness Analysis System

Liu et al. dan en [Liu et al., 2000] una segunda aproximación al problema de medir lo inesperado de una regla de acuerdo a las creencias del usuario.

Este trabajo también utiliza el sistema de IAS (Interestingness Analysis System) visto en la sección anterior y que podemos observar en de la Figura 2.1(d). Para que los usuarios especifiquen sus creencias utilizan un lenguaje no muy complejo que se basa en la representación del conocimiento mediante relaciones asociativas con los items que aparecen en la base de datos, y su sintaxis tiene la misma forma que una regla de asociación. El lenguaje permite tres grados de precisión para el conocimiento:

- impresiones generales (GI), están dadas por conjuntos de items que el usuario cree que están asociados entre sí pero no sabe de qué forma;
- conceptos razonablemente precisos (RPC), vienen dados por reglas que el usuario cree que podrán cumplirse; y
- conocimiento totalmente preciso (PK), es un conjunto de reglas que el usuario especifica de forma precisa.

Los primeros dos tipos representan el conocimiento con vaguedad y poco preciso que posee el usuario, mientras que el último tipo se refiere al conocimiento preciso.

El lenguaje que los autores utilizan también hace uso de la idea de jerarquía entre clases y los conceptos que pertenecen a las clases, pudiendo formar nuevas clases en cualquier momento del proceso según las necesidades del usuario.

Después de analizar el conocimiento sobre el dominio del usuario, el sistema lo utiliza para analizar las reglas obtenidas.

Sea U el conjunto de creencias del usuario y A el conjunto de las reglas de asociación obtenidas en el proceso de DM. La medida que utilizan los autores para ordenar las reglas depende del concepto de regla interesante que tengamos:

- *Reglas conformes.* Una regla $A_i \in A$ es conforme a una regla $U_j \in U$ del conocimiento del usuario si el antecedente y el consecuente de ambas reglas no difieren mucho.
- *Reglas con consecuente inesperado.* Una regla $A_i \in A$ tiene el consecuente inesperado con respecto a $U_j \in U$ si los antecedentes son similares pero no lo son los consecuentes.

- *Reglas con antecedente inesperado.* Es similar al caso anterior pero siendo en este caso los antecedentes muy distintos. Este tipo de reglas son muy útiles para el usuario debido a que el resultado que éste esperaba se consigue con distintos antecedentes de lo esperado.

- *Ambos lados de la regla inesperados.* Una regla $A_i \in A$ tiene ambos lados inesperados con respecto a $U_j \in U$ si los antecedentes y los consecuentes son inesperados. Estas reglas dirán al usuario que hay asociaciones que él no tuvo en cuenta o no conocía al especificar el conjunto de creencias inicial.

Una vez calculadas las medidas asociadas a estos cuatro tipos de reglas podremos ordenarlas de forma descendente respecto del conocimiento del usuario dado por U.

Trabajo 2.3. Item-Relatedness

Shekar y Natarajan basan el estudio de la medida de interés subjetiva tratando de definir un aspecto de "unexpectedness" que definen como "relación entre atributos" (item relatedness) [Shekar and Natarajan, 2004]. Para ello consideran el caso de las transacciones de la bolsa de la compra. En este contexto puede ocurrir que un mismo item pertenezca a varias categorías debido a que tenga características similares a otros items o funcionalidad parecida. El usuario espera que los items muy relacionados se compren juntos en una transacción. Por ejemplo, es normal que el que compre pan de molde también compre mantequilla. Pero es raro que se compren juntos pan de molde y bolígrafos. Luego este último caso si ocurriese, sería interesante porque contiene dos items no relacionados o muy poco relacionados. Así podemos mantener una relación inversa entre interés e items relacionados.

El proceso general que llevan a cabo los autores para identificar las reglas interesantes se basa en las siguientes apreciaciones:

- En primer lugar se obtienen las reglas de asociación de la base de datos usando los umbrales impuestos por el usuario de mínimo soporte y mínima confianza.

- Después, las creencias y expectativas del usuario son expresadas en forma de taxonomía[2] difusa, donde los nodos hoja muy próximos representan items que tienen similar funcionalidad.

[2]Una taxonomía es un árbol donde los nodos son conceptos y las ramas representan relaciones entre los nodos-padre y los nodos-hijo.

- Las propiedades estructurales de la taxonomía difusa se utilizan para calcular la relación entre varios items de una regla de asociación.
- El interés de la regla se calcula usando la estructura de la regla y las relaciones entre cada par de items en ella.
- El último paso es ordenar las reglas obtenidas dependiendo del grado de interés asociado.

Si observamos los dos enfoques anteriores, podemos identificarlos con la Figura 2.1(d), sin embargo este enfoque se ajusta al esquema representado en la Figura 2.3.

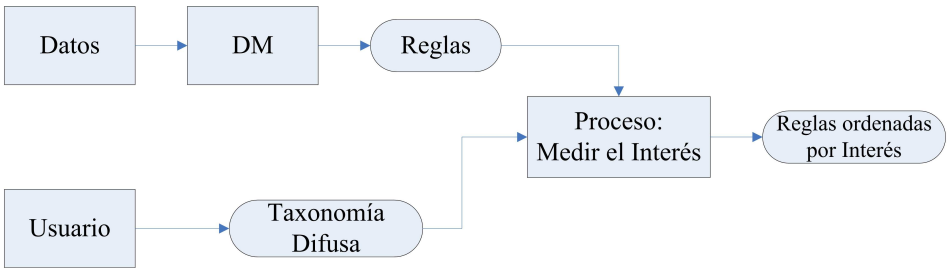

Figura 2.3: Aproximación para identificar patrones interesantes

Una vez introducido el concepto básico de medida de interés y después del breve recorrido por las distintas propuestas en este campo, nos centraremos en las propiedades que debería cumplir una buena medida objetiva de interés en la próxima sección.

2.1.4. Principios a Cumplir por las Medidas de Interés

En los apartados anteriores hemos presentado las distintas propuestas para medir el interés de una regla de asociación. Para establecer qué medida es la adecuada en cada situación, destacan varios trabajos que proponen unos requisitos mínimos que debería cumplir toda medida de interés. Estos requisitos, también llamados axiomas o principios, varían dependiendo de si evaluamos una medida para clasificación, una subjetiva o una objetiva. Entre los trabajos que llevan a cabo este cometido se encuentran [Piatetsky-Shapiro, 1991], [Major and Mangano, 1995], [Kamber and Shinghal, 1996], [Freitas, 1998] y [Hilderman and Hamilton, 2001].

Para evaluar las medidas de interés objetivas la propuesta hecha por Piatetsky-Shapiro [Piatetsky-Shapiro, 1991] junto con una extensión hecha por Major y Mangano [Major and Mangano, 1995] es la más aceptada por el resto de investigadores.

A continuación, veremos en qué consisten las distintas propuestas, las analizaremos y estudiaremos en profundidad y además discutiremos sobre la conveniencia de añadir dos nuevos principios a los tres ya presentados por Piatetsky-Shapiro. Más adelante en la sección 2.1.6 retomaremos esta discusión para justificar el uso de dichos principios (los tres de Piatetsky-Shapiro junto con los dos nuevos propuestos por nosotros) mediante el modelo lógico. De esta forma, estableceremos una relación directa entre los principios y algunas clases de cuantificadores-4ft. Para acabar la sección, probaremos que el factor de certeza [Berzal et al., 2002] cumple todos los principios usando para ello el modelo.

Propiedades Propuestas con Anterioridad

En la literatura sobre reglas de asociación podemos encontrar varias propuestas para enumerar cuáles son las propiedades deseables que debería tener una medida de interés. En particular, dichas propuestas suelen centrarse en el estudio de las propiedades de cualquier medida de interés objetiva.

El primer trabajo en esta línea fue el propuesto por Piatetsky-Shapiro en 1991 donde introdujeron tres principios que permiten establecer el comportamiento deseable para cualquier medida de interés a la que nombraremos por RI (Rule Interestingness) [Piatetsky-Shapiro, 1991].

Sea N el número total de transacciones que posee una base de datos D. Consideremos la regla de asociación $\varphi \rightarrow \psi$. Notaremos por $|\varphi|, |\psi|$ y $|\varphi \wedge \psi|$ al número de transacciones que contienen o satisfacen φ, ψ y $\varphi \wedge \psi$ respectivamente. Entonces las propiedades dadas por Piatetsky-Shapiro pueden enumerarse como sigue:

PS-1. $RI = 0$ si $|\varphi \wedge \psi| = |\varphi| |\psi| / N$. Si φ y ψ son estadísticamente independientes, entonces la regla no es interesante.

PS-2. RI es monótona creciente en $|\varphi \wedge \psi|$ si el resto de parámetros están fijos;

PS-3. RI es monótona decreciente en $|\varphi|$ (y en $|\psi|$) si el resto de parámetros están fijos.

Donde el conjunto de parámetros es justamente $\{|\varphi|, |\psi|, |\varphi \wedge \psi|\}$.

El primer principio propuesto por Piatetsky-Shapiro [Piatetsky-Shapiro, 1991] puede traducirse en que toda medida debe tener un valor significativo que indique cuándo hay independencia estadística entre el antecedente y el consecuente de la regla, ya que si esto ocurre basta hacer una traslación de dicha medida para que el valor 0 indique dicha independencia.

Además, el cumplimiento del principio PS-1 induce varias consecuencias: (1) RI es positiva si $|\varphi \wedge \psi| > |\varphi|\,|\psi|\,/N$, es decir, φ depende positivamente de ψ; y (2) RI es negativa si $|\varphi \wedge \psi| < |\varphi|\,|\psi|\,/N$, es decir, φ depende negativamente de ψ. Esto influye en la dirección de la regla de asociación, si φ depende positivamente de ψ entonces tenemos evidencia de que la regla $\varphi \rightarrow \psi$ es interesante, y si φ depende negativamente de ψ entonces tenemos evidencia de que la regla $\psi \rightarrow \varphi$ es interesante. Estos principios tienen otras implicaciones interesantes detalladas también en [Piatetsky-Shapiro, 1991].

Una de las medidas de interés que cumple todas estas propiedades es la propuesta por Piatetsky-Shapiro en [Piatetsky-Shapiro, 1991]:

$$RI = |\varphi \wedge \psi| - |\varphi|\,|\psi|\,/N \tag{2.1}$$

aunque hay otras muchas medidas que también satisfacen dichos principios (ver por ejemplo [Tan et al., 2002]).

En [Major and Mangano, 1995], Major y Mangano propusieron un cuarto principio complementario a los tres dados por Piatetsky-Shapiro pero en un contexto particular. Ellos utilizan la medida de interés para definir una relación de orden entre reglas. Cada regla tiene una medida asociada que viene representada por un punto en un espacio bidimensional donde las dos dimensiones son la cobertura (el soporte del antecedente) y el factor de confianza de la regla. Dicho principio puede enunciarse como sigue:

M-4. RI es monótona creciente en $|\varphi|$ si el factor de confianza está fijo y supera un umbral mínimo impuesto por el usuario.

Muchos autores han considerado dicho principio como una cuarta propiedad que debería cumplir toda medida de interés complementando las tres propuestas por Piatetsky-Shapiro.

Kamber y Shingal propusieron un quinto principio orientado a extender los cuatro anteriores pero en el contexto de reglas características, un tipo particular de reglas que no estudiaremos aquí [Kamber and Shinghal, 1996].

Freitas criticó en general la forma de formular los distintos principios que hemos comentado, argumentando que dicha formulación requiere que algunos factores no cambien, y por tanto, propone un conjunto de factores que varían dependiendo del tipo de medida de interés tomada [Freitas, 1999]. Dependiendo del valor de dichos factores se considerará que la medida es buena o no. Pero la mayoría de los factores que proponen están asociados a un tipo específico de reglas de asociación, en este caso, a las reglas de clasificación.

Hay otros trabajos que describen un conjunto de axiomas que debería cumplir toda medida de interés pero solamente son válidos para unos tipos específicos de reglas de asociación. En particular Hilderman y Hamilton describen un conjunto de cinco principios que debería cumplir cualquier medida de interés usada para ordenar resúmenes según su importancia generados desde la misma base de datos [Hilderman and Hamilton, 2001]. Tan et al. estudian distintas propiedades que puede cumplir una medida para poder distinguir las que sean más interesantes atendiendo a si cumplen o no dichas propiedades [Tan et al., 2002], como por ejemplo: simetría de la medida bajo permutación de una variable, variación a cambios de escala en los datos, etc.

Podemos observar que el cuarto principio que ha sido considerado hasta ahora, propuesto por Major y Mangano, es un poco difícil de formular y de extraer su significado. El primer problema que nos encontramos es el uso de un umbral en la definición, que puede ser útil para ordenar la importancia de las reglas y para desechar desde el principio aquellas que no sean interesantes. El problema es que en dicha situación puede parecer una propiedad deseable, pero en principio no dice nada concreto sobre la medida de interés. Además muchos autores lo han utilizado como cuarto principio porque no se podía obtener de los tres anteriores.

Si observamos con más detenimiento la definición de M-4 podemos deducir que la medida de interés que la cumpla, necesariamente es una medida no acotada. La siguiente proposición lo justifica [Delgado et al., 2010b].

Proposición 2.1. *Sea RI una medida de interés para reglas de asociación que satisface el principio M-4. Entonces RI no puede estar acotada.*

Demostración. Lo probaremos usando el contrarrecíproco.

Supongamos que RI está acotada. Como RI verifica el principio M-4 entonces RI es creciente en $|\varphi|$, es decir, RI crece cuando aumenta $|\varphi|$. Puesto que hemos supuesto que RI está acotada, supongamos que alcanza su máximo valor para una regla a la que llamaremos $r_1 = \varphi \rightarrow \psi$ en la base de datos D_1. Si tomamos la

base de datos D_2 que consiste en las transacciones de D_1 duplicadas, la regla r_1 también se encontrará en D_2 y tendrá el mismo valor de RI, que justamente es su máximo valor (ya que está acotada).

Esto contradice que RI cumpla M-4, ya que dicha medida debería crecer puesto que $|\varphi|$ se ha duplicado.

■

En la siguiente sección propondremos dos nuevos principios que complementan los tres propuestos por Piatetsky-Shapiro para poder distinguir las medidas de interés que a nuestro parecer consideraremos 'buenasó con un buen comportamiento.

2.1.5. Nuevos Principios para una Buena Medida de Interés

Hasta ahora hemos analizado los distintos principios propuestos en la literatura, y las propiedades recogidas en los principios PS-1, PS-2, PS-3 que son a nuestro juicio deseables para cualquier medida de interés. Pero son insuficientes, puesto que hay medidas como LIFT, Odds Ratio (α) o el Coeficiente de correlación (R) de la Tabla 2.1 que cumplen los principios PS-1, PS-2 y PS-3 y no están acotadas o no cumplen propiedades como las expuestas en [Tan et al., 2002] que pueden ser importantes dependiendo del tipo de regla de asociación que queramos obtener.

Por esta razón queremos añadir dos nuevos principios para que complementen a los tres anteriores [Delgado et al., 2010b] basándonos en la fuerza de la regla de asociación y en que esta debe estar acotada para poder imponer un umbral mínimo del cumplimiento de la misma. Los dos nuevos principios los notaremos por O-4 y O-5 y son los siguientes:

O-4. RI deberá estar acotada y ser menor o igual que 1.

O-5. RI deberá cumplir los cuatro principios anteriores para el contrarrecíproco de la regla también.

El cuarto principio O-4 está motivado por el establecimiento de umbrales para discernir entre la reglas interesantes y las que no lo son en absoluto. Un ejemplo son las reglas *fuertes* que son aquellas que satisfacen los umbrales de mínimo soporte y confianza. Este principio anula la suposición del principio M-4 debido a la proposición 2.1. Además imponemos que el máximo valor de RI sea 1. Si esto no ocurriese, puesto que la medida está acotada, bastaría normalizarla dividiendo por el máximo valor que pueda alcanzar.

El principio O-5 impone que cualquier medida de interés tenga el mismo comportamiento para una regla y su contrarrecíproca. Esta imposición está motivada por la propiedad lógica que establece la equivalencia entre una implicación y su contrarrecíproca ($\varphi \rightarrow \psi \equiv \neg\psi \rightarrow \neg\varphi$).

Además es una propiedad intuitiva a la que estamos predispuestos de manera inconsciente. Cuando tenemos dos bases de datos con el mismo número de transacciones satisfaciendo $\varphi \wedge \psi$, presentamos una mejor predisposición a aquella que tenga el mayor número de transacciones conteniendo $\neg\varphi \wedge \neg\psi$. Esto indica una relación más cercana entre el antecedente y el consecuente de la regla. Por tanto, tener una mayor evidencia en los datos del cumplimiento de una regla y su contrarrecíproca nos da mayor fiabilidad sobre la validez de la regla.

2.1.6. Relación entre los Principios y las Clases de Cuantificadores

Este apartado establece una conexión directa entre los cinco principios propuestos (PS-1, PS-2, PS-3, O-4 y O-5) y algunas de las clases de cuantificadores-4ft vistos en la sección 1.2. Esta conexión nos dará un método sencillo para demostrar cuándo un cuantificador-4ft cumple los anteriores principios.

Formulación de los Principios usando el Modelo Lógico

En primer lugar analizaremos los cinco principios usando el modelo lógico.

PS-1. $RI = 0$ si φ y ψ son estadísticamente independientes, es decir, si $a(a+b+c+d) = (a+b)(a+c)$. Pero esta condición puede simplificarse:

$$\begin{gathered} a(a+b+c+d) = (a+b)(a+c) \Leftrightarrow \\ a^2 + ab + ac + ad = a^2 + ac + ba + bc \Leftrightarrow \\ ad = bc. \end{gathered}$$

Por tanto, es razonable cambiar la condición de independencia por el cumplimiento de la igualdad $ad = bc$. Equivalentemente, $RI = 0$ si $ad - bc = 0$. Con esta notación podemos inferir dos interesantes propiedades: (1) si $ad - bc > 0$, entonces φ está positivamente asociada a ψ; y (2) si $ad - bc < 0$, entonces φ está negativamente asociada a ψ. Resumiendo, este principio nos dice que si $ad = bc$ entonces la regla no es interesante y por tanto la medida RI (en este caso el valor del cuantificador-4ft) debe ser cero.

PS-2. RI es monótona creciente en $a = |\varphi \wedge \psi|$ cuando $a + c$ y $a + b$ permanecen fijos (ya que el resto de parámetros que se tienen en consideración son $|\varphi|$ y $|\psi|$). Con una formulación matemática, podemos reescribir este principio de la siguiente forma:

$$\begin{aligned} &\text{Si } a_1 \leq a_2,\ a_1 + c_1 = a_2 + c_2 \text{ y } a_1 + b_1 = a_2 + b_2 \\ &\Rightarrow RI(a_1, b_1, c_1, d) \leq RI(a_2, b_2, c_2, d) \end{aligned}$$

PS-3a. RI es monótona decreciente en $a + b$ cuando a y $a + c$ permanecen fijos. Es decir, si a y $a + c$ permanecen constantes (equivalentemente, a y c no varían), la regla es menos interesante cuando b aumenta. Por tanto, la regla será más interesante cuando b disminuya y a, c permanezcan invariantes. Formalmente:

$$\begin{aligned} &\text{Si } b_1 \geq b_2,\ a_1 = a_2 \text{ y } c_1 = c_2 \\ &\Rightarrow RI(a_1, b_1, c_1, d) \leq RI(a_2, b_2, c_2, d) \end{aligned}$$

PS-3b. RI es monótona decreciente en $a + c$ cuando a y $a + b$ permanecen fijos. Es decir, si a y $a + b$ son constantes (equivalentemente, a y b no varían), la regla será menos interesante cuando c aumente. Por tanto, la regla será más interesante cuando c disminuya y a, b permanezcan invariantes. Matemáticamente:

$$\begin{aligned} &\text{Si } c_1 \geq c_2,\ a_1 = a_2 \text{ y } b_1 = b_2 \\ &\Rightarrow RI(a_1, b_1, c_1, d) \leq RI(a_2, b_2, c_2, d) \end{aligned}$$

O-4. El cuarto principio propuesto puede traducirse como: $\approx (a, b, c, d) \leq 1$ para todos a, b, c, d donde $n = a + b + c + d$, y $\approx (a, 0, 0, d) = 1$ para cualesquiera a, d donde $a + d = n$. Esto significa que la regla es muy interesante si todas las transacciones cumpliendo φ también cumplen ψ y aquellas satisfaciendo $\neg\varphi$ también satisfacen $\neg\psi$.

O-5. Recordemos que el principio O-5 se define en términos del cumplimiento de los cuatro principios anteriores pero para el contrarrecíproco de la regla. El anterior análisis se puede repetir de forma análoga para este caso.

O-5.1. $\neg\psi$ y $\neg\varphi$ son estadísticamente independientes si $d(a+b+c+d) = (b+d)(c+d)$, i.e. $RI = 0$ si φ y ψ son estadísticamente independientes, puesto que

$d(a+b+c+d) = (b+d)(c+d)$ es equivalente a $ad = bc$:

$$d(a+b+c+d) = (b+d)(c+d) \Leftrightarrow$$
$$ad + db + dc + d^2 = bc + bd + dc + d^2 \Leftrightarrow$$
$$ad = bc.$$

Luego, $\neg\psi$ y $\neg\varphi$ son estadísticamente independientes si y sólo si φ y ψ son estadísticamente independientes. Más aún, si $ad = bc$ entonces $\approx (d, b, c, a) = 0$ debe cumplirse, y es equivalente a imponer que si $ad = bc$ entonces $\approx (a, b, c, d) = 0$.

O-5.2. La regla es más interesante cuando d aumenta y $d + b$, $d + c$ permanecen constantes. Formalmente:

$$\text{Si } d_1 \leq d_2,\ d_1 + b_1 = d_2 + b_2 \text{ y } d_1 + c_1 = d_2 + c_2$$
$$\Rightarrow RI(a, b_1, c_1, d_1) \leq RI(a, b_2, c_2, d_2)$$

O-5.3a. La regla es más interesante cuando c disminuye y d, b no varían. Matemáticamente:

$$\text{Si } c_1 \geq c_2,\ d_1 = d_2 \text{ y } b_1 = b_2$$
$$\Rightarrow RI(a, b_1, c_1, d_1) \leq RI(a, b_2, c_2, d_2)$$

O-5.3b. La regla será más interesante cuando b disminuya y d, c permanecen constantes. Formalmente:

$$\text{Si } b_1 \geq b_2,\ d_1 = d_2 \text{ y } c_1 = c_2$$
$$\Rightarrow RI(a, b_1, c_1, d_1) \leq RI(a, b_2, c_2, d_2)$$

O-5.4. El cumplimiento del cuarto principio para la regla $\neg\psi \rightarrow \neg\varphi$ puede traducirse a que $\approx (d, b, c, a) \leq 1$ para todos a, b, c, d donde $n = a+b+c+d$, y $\approx (d, 0, 0, a) = 1$ para cualesquiera a, d con $n = a+d$. Estas dos restricciones son equivalentes a las condiciones de acotación impuestas para la regla $\varphi \rightarrow \psi$ ya que a y d juegan el mismo papel.

Relación entre los Principios y las Clases de Cuantificadores-4ft

Usando esta formulación veremos qué relación guardan los principios y algunas de las clases de cuantificadores-4ft presentadas en la sección 1.2.2.

Proposición 2.2. *Sea RI una medida de interés para reglas de asociación. Entonces,*

(I) *RI verifica PS-1 si, y sólo si, el cuantificador-4ft asociado a dicha medida cumple que* $\approx (a,b,c,d) = 0$ *cuando* $ad = bc$. *Las consecuencias del cumplimiento de PS-1 enumeradas como (1) y (2) se cumplen también cuando* $\approx$ *es un cuantificador-4ft comparativo.*

(II) *RI verifica PS-1 para el contrarrecíproco de la regla si, y sólo si, el cuantificador-4ft asociado cumple que* $\approx (d,b,c,a) = 0$ *cuando* $ad = bc$. *Las consecuencias de PS-1 para el contrarrecíproco de la regla se verifican si* $\approx$ *es comparativo.*

Demostración. (I) La primera afirmación es trivial utilizando el anterior análisis:

$$RI = 0 \text{ si } |\varphi \wedge \psi| = |\varphi|\,|\psi|\,/N \Longleftrightarrow \approx (a,b,c,d) = 0 \text{ si } ad = bc.$$

Si el cuantificador-4ft $\approx$ asociado a la regla $\varphi \approx \psi$ es *comparativo*, entonces $Val(\varphi \approx \psi) = true$ implica que $ad > bc$; i.e. φ y ψ asociados positivamente (RI positiva) implica que $ad > bc$. Al contrario, $Val(\varphi \approx \psi) = false$ implicará que $ad < bc$.

(II) Análogo al caso *(I)*.

■

Proposición 2.3. *Sea RI una medida de interés para reglas de asociación. Entonces*

(I) *PS-2 y PS-3 se cumplen si, y sólo si, el cuantificador-4ft asociado* $\approx$ *es doble implicacional.*

(II) *PS-2 y PS-3 se cumplen para la regla y su contrarrecíproca si, y sólo si el cuantificador-4ft* $\approx$ *asociado es un cuantificador-4ft equivalente.*

Demostración. El resultado es trivial a partir del anterior análisis (PS-2, PS-3a y PS-3b para *(I)* y O-5.2, O-5.3a y O-5.3b para *(II)*). ■

Proposición 2.4. *Sea RI una medida de interés para reglas de asociación. Entonces*

(I) *RI cumple O-4 si, y sólo si, el cuantificador-4ft asociado satisface:*

$$\approx (a,b,c,d) \leq 1 \qquad \forall\, a,b,c,d$$

donde $|D| = a+b+c+d$, *y*

$$\approx (a,0,0,d) = 1 \qquad \forall\, a,d \;\; \textit{tales que} \;\; a+d=n.$$

(II) *RI verifica O-4 para el contrarrecíproco de la regla si, y sólo si, el cuantificador asociado satisface:*

$$\approx (d, b, c, a) \leq 1 \qquad \forall\, a, b, c, d$$

con $|D| = a + b + c + d$, *y*

$$\approx (d, 0, 0, a) = 1 \qquad \forall\, a, d \;\; \textit{tales que} \;\; a + d = n.$$

Demostración. La demostración es trivial. ■

Corolario 2.1. *Una medida de interés RI cumplirá los principios PS-1, PS-2 y PS-3, si y sólo si, el cuantificador-4ft asociado* $\approx$ *verifica lo siguiente:*

- $\approx$ *es comparativo,*
- *si* $ad = bc$ *entonces* $\approx (a, b, c, d) = 0$, *y*
- $\approx$ *es doble implicacional.*

Corolario 2.2. *Una medida de interés RI cumplirá PS-1, PS-2, PS-3, O-4 y O-5 si, y sólo si, el cuantificador-4ft asociado* $\approx$ *verifica lo siguiente:*

- $\approx$ *es comparativo,*
- *si* $ad = bc$ *entonces* $\approx (a, b, c, d) = 0$,
- $\approx (a, b, c, d) \leq 1\ \forall\ a, b, c, d$ *donde* $n = a + b + c + d$,
- $\approx (a, 0, 0, d) = 1\ \forall a, d$ *con* $n = a + d$, *y*
- $\approx$ *es un cuantificador equivalente.*

Demostración. Es necesario tener en cuenta que *(I)* y *(II)* en las proposiciones 2.2 y 2.4 son equivalentes ya que el papel de a y d se puede intercambiar en ambos casos. ■

Esta nueva propuesta da una forma sencilla de probar si una medida de interés para reglas de asociación (o el cuantificador-4ft que se define a partir de dicha medida) cumple los tres primeros principios (desde PS-1 hasta PS-3) y los dos nuevos principios que hemos considerado. También queremos remarcar que los principios O-4 y O-5 propuestos por nosotros en el capítulo anterior sólo afectan al cumplimiento de algunas condiciones extra, siendo una de ellas que el cuantificador esté acotado, lo cual, se reduce a probar dos simples condiciones. La otra se

reduce a probar que el cuantificador sea equivalente en lugar de solamente doble implicacional.

A continuación veremos cómo el factor de certeza [Berzal et al., 2002] puede verse como un cuantificador-4ft. Después probaremos teóricamente que es una buena medida de interés puesto que cumple los cinco principios anteriores.

2.1.7. El Cuantificador-4ft Factor de Certeza

La teoría de los *factores de certeza* fue introducido por Shortliffe y Buchanan en [Shortliffe and Buchanan, 1975] y fue utilizada por Sánchez [Sánchez, 1999], [Berzal et al., 2002] como una alternativa a la confianza.

Definición 2.1. [Delgado et al., 2003a] El *factor de certeza*, FC, de una regla $\varphi \rightarrow \psi$ se define como

$$FC(\varphi \rightarrow \psi) = \begin{cases} \dfrac{\text{Conf}(\varphi \rightarrow \psi) - \text{sop}(\psi)}{1 - \text{sop}(\psi)} & \text{si } \text{Conf}(\varphi \rightarrow \psi) > \text{sop}(\psi) \\ \dfrac{\text{Conf}(\varphi \rightarrow \psi) - \text{sop}(\psi)}{\text{sop}(\psi)} & \text{si } \text{Conf}(\varphi \rightarrow \psi) < \text{sop}(\psi) \end{cases} \tag{2.2}$$

imponiendo que si $\text{sop}(\psi) = 1$ entonces $FC(\varphi \rightarrow \psi) = 1$ y si $\text{sop}(\psi) = 0$ entonces $FC(\varphi \rightarrow \psi) = -1$.

El factor de certeza toma valores entre $[-1, 1]$. Es positivo si φ y ψ están asociados positivamente, es 0 cuando hay independencia y, es negativo si están asociados negativamente. El factor de certeza puede interpretarse como una medida de la variación de la probabilidad de que el consecuente ψ esté en una transacción cuando consideramos sólo aquellas transacciones en las que se cumple φ. De forma más específica, un FC positivo medirá el crecimiento en la probabilidad de que ψ no esté en una transacción sabiendo que φ sí lo está. Para los valores negativos tiene una interpretación similar.

Además el factor de certeza reduce el número de reglas obtenidas siendo éstas más fuertes que las obtenidas con la confianza (ver [Delgado et al., 2003a] y [Berzal et al., 2001] para una descripción más detallada del problema). El uso del duo *soporte/FC* soluciona también un problema esencial en la falta de semántica de la probabilidad condicional, y en consecuencia, de la semántica de la confianza. El factor de certeza provee un significado útil para medir la validez de una regla de asociación y se utilizará para la extracción de reglas *muy fuertes* como veremos en la siguiente sección.

A continuación definiremos el cuantificador-4ft factor de certeza y después estudiaremos sus propiedades usando el modelo lógico.

En primer lugar necesitaremos definir el soporte de un itemsets en términos del modelo. Para ello extenderemos el concepto de soporte para itemsets en lugar de para reglas. Si φ y ψ son itemsets y $4ft(\varphi, \psi, D) = \langle a, b, c, d\rangle$ es su tabla-4ft asociada, sus soportes vendrán dados por

$$\mathrm{sop}(\varphi) = \frac{a+b}{a+b+c+d}, \quad \mathrm{sop}(\psi) = \frac{a+c}{a+b+c+d}.$$

Ahora sí podemos definir el cuantificador-4ft *factor de certeza*, $\approx_{FC}$, de una regla de asociación $\varphi \approx \psi$ como

$$\approx_{FC}(a,b,c,d) = \begin{cases} \dfrac{\Rightarrow_I (a,b) - \mathrm{sop}(\psi)}{1 - \mathrm{sop}(\psi)} & \text{si } \Rightarrow_I (a,b) > \mathrm{sop}(\psi) \\ \dfrac{\Rightarrow_I (a,b) - \mathrm{sop}(\psi)}{\mathrm{sop}(\psi)} & \text{si } \Rightarrow_I (a,b) < \mathrm{sop}(\psi) \\ 0 & \text{en otro caso.} \end{cases} \tag{2.3}$$

Recordemos que imponer que $\Rightarrow_I (a,b) < \mathrm{sop}(\psi)$ es equivalente a

$$\frac{a}{a+b} < \frac{a+c}{a+b+c+d}$$

y por tanto a $ad < bc$. Análogamente $\Rightarrow_I (a,b) > \mathrm{sop}(\psi)$ es equivalente a $ad > bc$.

El cuantificador-4ft $\approx_{FC}$ induce los predicados lógicos asociados a la regla de asociación $\varphi \approx_{FC} \psi$ como veremos a continuación.

Sean $0 < p < 1$ y $0 < Base < a+b+c+d$ dos umbrales predefinidos por el usuario, entonces se cumple lo siguiente:

- Si $ad > bc$, i.e. φ y ψ están asociados positivamente, entonces

$$Val(\varphi \approx_{FC} \psi) = true \iff$$
$$\approx_{FC}(a,b,c,d) \geq p \quad \wedge \quad a \geq Base \quad \wedge \quad d \geq Base.$$

- Si $ad = bc$, entonces φ y ψ son estadísticamente independientes $\iff$ $\approx_{FC}(a,b,c,d) = 0$.

- Si $ad < bc$, entonces φ y ψ están asociados negativamente $\iff$ $\approx_{FC}(a,b,c,d) < 0$.

El factor de certeza da más información que la confianza sobre los itemsets involucrados en la regla, ya que también permite ver si están asociados negativamente. Observemos que las dos primeras propiedades prueban que el factor de certeza cumple la propiedad PS-1 dada por Piatetsky-Shapiro. Seguiremos probando que el $\approx_{FC}$ es un cuantificador equivalente.

Proposición 2.5. *Sea* $\approx_{FC}\ (a,b,c,d)$ *el cuantificador-4ft factor de certeza definido por la ecuación (2.3). Entonces* $\approx_{FC}\ (a,b,c,d) \in E$*, i.e.* $\approx_{FC}\ (a,b,c,d)$ *es un cuantificador-4ft equivalente.*

Demostración. Tenemos que probar que si se cumple

$$a' \geq a \quad \wedge \quad b' \leq b \quad \wedge \quad c' \leq c \quad \wedge \quad d' \geq d$$

entonces debe cumplirse que

$$\approx_{FC}\ (a',b',c',d') \geq \approx_{FC}\ (a,b,c,d).$$

Para ello, es fácil ver que la definición dada mediante la ecuación (2.3) es equivalente a la siguiente

$$\approx_{FC}\ (a,b,c,d) = \begin{cases} \dfrac{ad-bc}{(a+b)(b+d)} & \text{si } ad > bc \\ 0 & \text{si } ad = bc \\ \dfrac{ad-bc}{(a+b)(a+c)} & \text{si } ad < bc. \end{cases} \tag{2.4}$$

Supongamos que

$$a' = a+k, \quad b' = b-l, \quad c' = c-s, \quad d' = d+p$$

donde $k,l,s,p \geq 0$. Dividiremos la demostración en tres casos distintos:

Caso 1. Si $ad > bc$ entonces la desigualdad $a'd' > b'c'$ también se cumple. Bajo esta suposición deberemos probar que

$$\approx_{FC}\ (a+k,b-l,c-s,d+p) - \approx_{FC}\ (a,b,c,d) \geq 0.$$

En este caso,

$$\frac{bc-ad}{(a+b)(b+d)} + \frac{(a+k)(d+p)-(b-l)(c-s)}{(a+k+b-l)(b-l+d+p)} = \frac{(N1)}{(D1)} \geq 0$$

donde

$$\begin{aligned} (N1) &= (bc-ad)(a+k+b-l)(b-l+d+p) + \\ &\quad + [(a+k)(d+p)-(b-l)(c-s)] * (a+b)(b+d) \\ (D1) &= (a+b)(b+d)(a+k+b-l)(b-l+d+p) \end{aligned}$$

Es suficiente probar que $(N1) \geq 0$ porque $(D1)$ es siempre positivo ya que $b-l \geq 0$.

Podemos simplificar $(N1)$ de la siguiente forma

$$\begin{aligned}(N1) = {} & adl(a+c+d) + b^2(ck+dk+ap+cp+kp) + \\ & + b(dk(c+d) + adl + cl^2 + pa^2 + acp + kp(a+c+d)) + \\ & + l(ad-bc)(b+k+p) + s(b-l)(ab+b^2+ad+bd)\end{aligned}$$

Así es fácil ver que $(N1)$ es ≥ 0 usando que $ad > bc$ y que $b - l \geq 0$.

Caso 2. Si $ad < bc$ y $a'd' < b'c'$, deberemos probar que se cumple la siguiente desigualdad:

$$\frac{bc-ad}{(a+b)(a+c)} + \frac{(a+k)(d+p)-(b-l)(c-s)}{(a+k+b-l)(a+k+c-s)} = \frac{(N2)}{(D2)} \geq 0$$

donde

$$\begin{aligned}(N2) = {} & (bc-ad)(a+k+b-l)(a+k+c-s) \\ & + [(a+k)(d+p)-(b-l)(c-s)] * (a+b)(a+c) \\ (D2) = {} & (a+b)(a+c)(a+k+b-l)(a+k+c-s)\end{aligned}$$

Luego es suficiente probar que $(N2) \geq 0$ porque $(D2)$ es otra vez positivo. Simplificando $(N2)$ así:

$$\begin{aligned}(N2) = {} & a^3p + bck(d+p) + a^2(p(c+k) + d(l+s) + b(p+s)) \\ & + ab(kp+ck+cp+ds) + ak(cp+dl+ds) + (bc-ad)(k^2+ak) + \\ & + (b-l)(bck+sba) + (c-s)(cbk+lad+a^2l+cla)\end{aligned}$$

es fácil ver que $(N2) \geq 0$ puesto que $bc - ad > 0$, $b - l \geq 0$ y $c - s \geq 0$.

Caso 3. Si $ad < bc$ y $a'd' > b'c'$, tenemos que probar que se satisface

$$\frac{bc-ad}{(a+b)(a+c)} + \frac{a'd'-b'c'}{(a'+b')(b'+d')} \geq 0$$

pero esto es inmediato debido a que se cumplen $ad < bc$ y $a'd' > b'c'$.

■

Gracias a esta proposición podemos reescribir el cuantificador factor de certeza como $\equiv_{FC}$ ya que es un cuantificador equivalente.

Además, $\equiv_{FC}$ depende de d, luego no es implicacional. La siguiente pregunta que nos podríamos hacer es si es un cuantificador Σ-equivalente, pero mediante un contraejemplo podemos ver que no es así. El cuantificador $\equiv_{FC}$ sería Σ-equivalente si siempre que

$$a' + d' \geq a + d \quad \wedge \quad b' + c' \leq b + c$$

se cumple la desigualdad

$$\equiv_{FC}(a', b', c', d') \geq \equiv_{FC}(a, b, c, d).$$

Si tomamos $(a, b, c, d) = (3, 1, 1, 2)$ y $(a', b', c', d') = (5, 1, 1, 1)$, es fácil comprobar que $a' + d' = 6 \geq 5 = a + d$ y $b' + c' = 2 = b + c$, pero

$$\equiv_{FC}(5, 1, 1, 1) = \frac{1}{3} \backsimeq 0.33 < 0.41 \backsimeq \frac{5}{12} = \equiv_{FC}(3, 1, 1, 2)$$

donde hemos usado la primera parte de la definición dada en la ecuación (2.3) ya que $a'd' = 5 > 1 = b'c'$ y $ad = 6 > 1 = bc$. Concluimos entonces que el factor de certeza no es un cuantificador Σ-equivalente.

La proposición 2.5 establece que el factor de certeza pertenece a la clase de los cuantificadores equivalentes. Además hemos visto que $\equiv_{FC}$ cumple el primer principio PS-1, está acotada por 1 y $\equiv_{FC}(a, 0, 0, d) = 1$, luego cumple los cinco principios propuestos para las medidas de interés (corolario 2.2).

Además el cuantificador factor de certeza mide la asociación positiva ($ad > bc$) y la negativa ($ad < bc$) entre el antecedente y el consecuente de la regla.

El análisis hecho en esta sección pone de manifiesto lo simple que es probar formalmente las propiedades de los cuantificadores-4ft y en consecuencia, las cinco propiedades propuestas (PS-1, PS-2, PS-3, O-4 y O-5) para las medidas de interés para reglas de asociación. El factor de certeza fue estudiado anteriormente en [Delgado et al., 2003a], [Delgado et al., 2003b] donde se prueba empíricamente que provee mejores resultados que la confianza, extrayendo además un conjunto menor de reglas, propiedad muy interesante cuando se trabaja con bases de datos de gran tamaño. Por tanto, en esta sección hemos dado una demostración formal que prueba el porqué de los buenos resultados que obteníamos en la práctica, puesto que el factor de certeza cumple los cinco principios que hemos propuesto para una 'buena' medida de interés.

A continuación definiremos en términos del modelo formal algunas propiedades y conceptos de interés para reglas de asociación que nos serán de utilidad en las próximas secciones.

2.1.8. Una Nueva Formulación para Reglas Fuertes y Muy Fuertes

El procedimiento general para calcular reglas de asociación consiste en encontrar aquellas reglas cuyo soporte exceda el mínimo impuesto por el usuario y

que el valor que mide la relación entre el antecedente y el consecuente sea lo mayor posible. La medida más famosa para medir dicha relación es la confianza, aunque hemos visto que no es la única, ni siquiera la mejor entre todas las propuestas. Pero al ser la más usada, se suele emplear el término *fuerte* para decir que el soporte y la confianza de la regla exceden los umbrales mínimos impuestos llamados *minsop* y *minconf* respectivamente.

Pero muchos autores han estudiado algunos de los inconvenientes del uso del par *soporte/confianza* para calcular reglas de asociación [Sánchez, 1999], [Brin et al., 1997], [Silverstein et al., 1998], [Berzal et al., 2002]. Para solucionar algunos de estos inconvenientes y para asegurarnos que las reglas descubiertas sean interesantes y precisas, en [Sánchez, 1999] y [Berzal et al., 2002] se propone el concepto de regla *muy fuerte* que veremos a continuación.

Definición 2.2. [Sánchez, 1999] Una regla de asociación $\varphi \rightarrow \psi$ es *muy fuerte* si las reglas $\varphi \rightarrow \psi$ y $\neg\psi \rightarrow \neg\varphi$ son fuertes.

Para calcular la precisión de las reglas de asociación nosotros usamos el factor de certeza en vez de la confianza. El factor de certeza tiene propiedades muy interesantes que podemos encontrar en [Sánchez, 1999] y que hemos justificado en la sección anterior. Además también se utiliza para extraer reglas muy fuertes como veremos a continuación.

Según la definición anterior, para que la regla de asociación $\varphi \rightarrow \psi$ sea muy fuerte usando el factor de certeza es necesario que se verifiquen las siguientes condiciones [Sánchez, 1999]:

1. $Sop(\varphi \rightarrow \psi) \geq minsop$,
2. $Sop(\neg\psi \rightarrow \neg\varphi) \geq minsop \Longleftrightarrow 1 - sop(\varphi) - sop(\psi) + Sop(\varphi \rightarrow \psi) \geq minsop$,
3. $FC(\varphi \rightarrow \psi) \geq minFC$, y
4. $FC(\neg\psi \rightarrow \neg\varphi) \geq minFC$.

Las dos últimas condiciones son equivalentes si $FC(\varphi \rightarrow \psi) > 0$. Usaremos estas condiciones para definir formalmente las dos posibilidades para acotar el valor de los cuantificadores-4ft a los que llamaremos un uso *fuerte* o *muy fuerte* de los cuantificadores-4ft.

Observemos que la tabla-4ft $\mathcal{M}^c$ asociada al contrarrecíproco de la regla $\varphi \approx \psi$ conocida $\mathcal{M} = \langle a, b, c, d \rangle$ es:

$\mathcal{M}^c$	$\neg\varphi$	φ
$\neg\psi$	d	b
ψ	c	a

Analizando las propiedades del factor de certeza usando este modelo lógico es fácil probar que las dos primeras condiciones son equivalentes a imponer que $a \geq Base$ y $d \geq Base$ respectivamente, donde $0 < Base < a+b+c+d$. Las dos últimas, nos dicen que dado un umbral llamémosle $0 < p < 1$, el valor del cuantificador-4ft para las reglas de asociación debe satisfacer que $\approx (a, b, c, d) \geq p$ y $\approx (d, b, c, a) \geq p$.

Definición 2.3. Un cuantificador-4ft $\approx$ se usa de un modo *fuerte* cuando se imponen las dos condiciones siguientes:

$$\approx (a, b, c, d) \geq p \quad \wedge \quad a \geq Base$$

donde $0 < p < 1$ y $0 < Base < a + b + c + d$. Además se usará de un modo *muy fuerte* cuando se imponen las siguientes condiciones:

$$\approx (a, b, c, d) \geq p \quad \wedge \quad \approx (d, b, c, a) \geq p \quad \wedge \quad a \geq Base \quad \wedge \quad d \geq Base.$$

Dependiendo de las imposiciones que establezcamos para el valor del cuantificador, estaremos buscando reglas fuertes o muy fuertes siguiendo la definición anterior. En particular, el cuantificador-4ft factor de certeza $\equiv_{FC}$ se usa siempre de un modo muy fuerte, esto es debido a que a y d son intercambiables al definir $\equiv_{FC}$.

Definición 2.4. Un cuantificador-4ft $\approx$ es *contrasimétrico* si

$$\approx (a, b, c, d) = \approx (d, b, c, a).$$

Usando las dos últimas definiciones (2.3 y 2.4) es sencillo probar que un cuantificador contrasimétrico se usará de un modo *muy fuerte* si satisface solamente las tres condiciones siguientes:

$$\approx (a, b, c, d) \geq p \quad \wedge \quad a \geq Base \quad \wedge \quad d \geq Base \tag{2.5}$$

donde $0 < p < 1$ y $0 < Base < a + b + c + d$.

Observemos que para que un cuantificador *contrasimétrico* cumpla los cinco principios (PS-1 hasta O-5) bastará con que satisfaga los cuatro primeros. Un ejemplo de cuantificador contrasimétrico es el factor de certeza $\equiv_{FC}$ pero sólo cuando $ad - bc \geq 0$, es decir cuando mide asociaciones positivas.

2.1.9. Conclusiones

El análisis desarrollado en esta sección nos muestra una conexión directa entre la clase a la que pertenece un cuantificador-4ft y las propiedades de la medida de interés que tiene asociada. Puesto que el uso de los cuantificadores-4ft, ya sea fuerte o muy fuerte, viene dado mediante su acotación por umbrales, hemos impuesto el principio O-4 que es una condición indispensable cuando se precisa restringir o acotar la búsqueda de reglas de asociación. El principio O-5 lleva implícita la idea de implicación que contiene el concepto de regla de asociación. Aunque debemos distinguir entre asociación y casualidad, que sería la propiedad asociada al concepto formal de implicación, es cierto que si una medida de interés mide el grado de asociación del contrarrecíproco de la regla, más seguros estaremos del vínculo existente entre dos itemsets.

Mediante la adición de estos principios a los tres de Piatetsky-Shapiro, obtenemos que comprobar el cumplimiento de todos ellos es (a grandes rasgos) lo mismo que probar que el cuantificador-4ft asociado es comparativo (distingue independencia), acotado y equivalente. Generalmente la comprobación de dichas propiedades es sencilla y simplemente requiere un poco de cálculo del tipo de la proposición 2.5, que será más complejo dependiendo de la medida en cuestión.

En la siguiente sección, continuaremos con el desarrollo del modelo formal para obtener una extensión consistente para la extracción de reglas de asociación difusas. Como consecuencia de la generalización del modelo, presentamos una forma sencilla de extender medidas de interés para validar reglas difusas.

2.2. Modelo para la Extracción de Reglas Difusas mediante Niveles de Restricción

El modelo GUHA proporciona una forma adecuada de representar la información contenida en una base de datos cuando fijamos dos itemsets mediante la llamada tabla-4ft, así como de las medidas utilizadas para medir la validez, utilidad, independencia, interés, etcétera entre ellos. Esta sección pretende ofrecer otra de las ventajas de utilizar este modelo para representar y evaluar reglas de asociación: su portabilidad y adaptabilidad a otros ambientes. En [Karban, 2003], [Karban et al., 2004], [Hájek et al., 2004] podemos ver algunos de los esfuerzos hechos para adaptar el modelo para extraer diversos tipos de reglas: reglas SDS (Sets Differs from Sets) y reglas de asociación condicionales.

En esta sección, desarrollaremos el modelo formal para representar reglas de asociación difusas. Esta extensión nos permitirá construir un método para extender las medidas de interés para reglas de asociación en medidas para reglas difusas [Delgado et al., 2008e], [Delgado et al., 2009]. Este método utilizará básicamente dos herramientas: el modelo lógico y el sistema de representación mediante niveles de restricción visto en la sección 1.3.

2.2.1. Generalización del Modelo para Reglas Difusas

Muchas de las medidas propuestas para validar las reglas de asociación crisp están definidas en términos de probabilidades. El soporte y la confianza son ejemplos de ello. El modelo de representación mediante niveles de restricción ofrece una buena herramienta para extender conceptos crisp al campo difuso, en particular, usando RL-probabilidades este tipo de medidas pueden ser fácilmente extendidas al caso difuso como presentamos a continuación.

Consideremos los conjuntos difusos $\tilde{\Gamma}_A$ y $\tilde{\Gamma}_B$ definidos en $\tilde{D}$ como $\tilde{\Gamma}_A(\tilde{\tau}) = \tilde{\tau}(A)$ y $\tilde{\Gamma}_B(\tilde{\tau}) = \tilde{\tau}(B)$ respectivamente[3] (ver definición 1.7). Para las medidas de interés que pueden ser expresadas en términos de probabilidad, podremos calcular las RL-representaciones asociadas a las RL-probabilidades. En nuestro caso, las RL-probabilidades serán $P(\tilde{\Gamma}_A)$ y $P(\tilde{\Gamma}_B)$ que notaremos por $P(\tilde{A})$ y $P(\tilde{B})$ respectivamente, y sus RL-representaciones por $(\Lambda_{P(\tilde{A})}, \rho_{P(\tilde{A})})$, $(\Lambda_{P(\tilde{B})}, \rho_{P(\tilde{B})})$. Del mismo modo, calcularemos las RL-probabilidades de la conjunción de $\tilde{\Gamma}_A$ y $\tilde{\Gamma}_B$, la RL-probabilidad condicionada, etc, todas ellas siguiendo las ecuaciones (1.20) y (1.23).

[3] A y B son conjuntos crisp de items disjuntos y $\tilde{\Gamma}_A$, $\tilde{\Gamma}_B$ son conjuntos difusos definidos en $\tilde{D}$.

Usando las *RL*-probabilidades podemos generalizar cualquier medida dada en términos de probabilidad como mostramos en las definiciones siguientes para los casos particulares de soporte y confianza.

Definición 2.5. (Soporte de un itemsets) Sea $A \subseteq I$ un itemsets y $(\Lambda_{P(\tilde{A})}, \rho_{P(\tilde{A})})$ la representación de la *RL*-probabilidad asociada al conjunto difuso $\tilde{\Gamma}_A$ en $\tilde{D}$. Entonces, el soporte de A en la base de datos difusa $\tilde{D}$ se define como

$$\text{sop}(A) = \sum_{\alpha_i \in \Lambda_{P(\tilde{A})}} (\alpha_i - \alpha_{i+1}) \left(\rho_{P(\tilde{A})}(\alpha_i)\right). \tag{2.6}$$

Siguiendo un razonamiento parecido definiremos el soporte y la confianza para una regla difusa $A \to B$.

Definición 2.6. (Soporte de $A \to B$) Sean $A, B \subseteq I$ dos itemsets disjuntos y sea $(\Lambda_{P(\tilde{A}\wedge\tilde{B})}, \rho_{P(\tilde{A}\wedge\tilde{B})})$, la *RL*-representación de la *RL*-probabilidad $P(\tilde{\Gamma}_A \wedge \tilde{\Gamma}_B)$ en $\tilde{D}$. Definimos el *soporte* de la regla difusa $A \to B$ en $\tilde{D}$ como

$$\text{Sop}(A \to B) = \sum_{\alpha_i \in \Lambda_{P(\tilde{A}\wedge\tilde{B})}} (\alpha_i - \alpha_{i+1}) \left(\rho_{P(\tilde{A}\wedge\tilde{B})}(\alpha_i)\right). \tag{2.7}$$

Definición 2.7. (Confianza de $A \to B$) Sean $A, B \subseteq I$ dos itemsets disjuntos y sea $(\Lambda_{P(\tilde{B}|\tilde{A})}, \rho_{p(\tilde{B}|\tilde{A})})$ la *RL*-representación de la *RL*-probabilidad $P(\tilde{\Gamma}_B|\tilde{\Gamma}_A)$ en $\tilde{D}$. Definimos la *confianza* de la regla difusa $A \to B$ en $\tilde{D}$ como

$$\text{Conf}(A \to B) = \sum_{\alpha_i \in \Lambda_{P(\tilde{B}|\tilde{A})}} (\alpha_i - \alpha_{i+1}) \left(\rho_{P(\tilde{B}|\tilde{A})}(\alpha_i)\right). \tag{2.8}$$

Queremos remarcar que en las anteriores definiciones $\Lambda_{P(\tilde{A}\wedge\tilde{B})}$ y $\Lambda_{P(\tilde{B}|\tilde{A})}$ son justamente $\Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}$.

Acabamos de ver cómo se puede generalizar de forma sencilla las medidas de interés que están en términos de probabilidades, pero hay otras medidas que no tienen esta propiedad. Para ellas, presentaremos un método por el que pueden ser generalizadas también al caso difuso. En este proceso usaremos que cualquier medida de interés utiliza las frecuencias entre las posibles combinaciones con los itemsets involucrados en la regla de asociación. Esta propiedad es la que utiliza el modelo formal que estamos estudiando. Mediante una adecuada extensión, obtenemos un nuevo modelo para reglas de asociación difusas basado en el anterior que será la pieza clave en el proceso de generalización de las medidas de interés.

Consideremos los conjuntos difusos $\tilde{\Gamma}_A$ y $\tilde{\Gamma}_B$ definidos en $\tilde{D}$ como $\tilde{\Gamma}_A(\tilde{\tau}) = \tilde{\tau}(A)$ y $\tilde{\Gamma}_B(\tilde{\tau}) = \tilde{\tau}(B)$ respectivamente. De acuerdo con lo expuesto en la sección 1.1.2

podemos obtener sus respectivas RL-representaciones que notaremos por $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}})$, $(\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$, definidas como se describe en las ecuaciones (1.20) y (1.23).

Los conjuntos de items: $A \wedge B, A \wedge \neg B, \neg A \wedge B, \neg A \wedge \neg B$ forman una partición de la base de datos $\tilde{D}$ (al igual que lo hacían $\varphi \wedge \psi, \varphi \wedge \neg\psi, \neg\varphi \wedge \psi$ y $\neg\varphi \wedge \neg\psi$ para el caso crisp). Del mismo modo, podemos considerar los conjuntos difusos asociados definidos en $\tilde{D}$ con sus respectivas RL-representaciones a las que llamaremos $(\Lambda_{\tilde{A}\wedge\tilde{B}}, \rho_{\tilde{A}\wedge\tilde{B}})$, $(\Lambda_{\tilde{A}\wedge\neg\tilde{B}}, \rho_{\tilde{A}\wedge\neg\tilde{B}})$, etc. Observemos que los RL-sets que aparecen contendrán los mismos niveles de restricción, es decir, todos ellos serán igual a $\Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}$.

Para cada $\alpha \in \Lambda_Y$, $\rho_Y(\alpha)$ es un conjunto crisp, luego podremos calcular su cardinal como viene siendo habitual, y lo notaremos por $|\rho_Y(\alpha)|$.

De esta forma para cada nivel de restricción $\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}$ definimos la tabla $4ft$ asociada $\mathcal{M}_{\alpha_i} = 4ft(\tilde{\Gamma}_A, \tilde{\Gamma}_B, \tilde{D}, \alpha_i)$ como

$\mathcal{M}_{\alpha_i}$	$\tilde{\Gamma}_B$	$\tilde{\Gamma}_{\neg B}$
$\tilde{\Gamma}_A$	a_i	b_i
$\tilde{\Gamma}_{\neg A}$	c_i	d_i

donde a_i, b_i, c_i y d_i serán enteros no negativos tales que $a_i = |\rho_{\tilde{A}\wedge\tilde{B}}(\alpha_i)|$, $b_i = |\rho_{\tilde{A}\wedge\neg\tilde{B}}(\alpha_i)|$ y análogamente con c_i y d_i.

Observemos que $\forall \alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}$ se tiene que

$$a_i + b_i + c_i + d_i = n = |\tilde{D}|. \tag{2.9}$$

Usando la representación que nos ofrece el modelo lógico para reglas difusas, es fácil generalizar cualquier tipo de medida de interés crisp al caso difuso, en particular, cualquier cuantificador $4ft$. A continuación lo comprobaremos para los casos particulares de soporte y confianza.

Definición 2.8. (Soporte de un itemsets) Sea $A \subseteq I$ un itemsets y $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}})$ la RL-representación asociada al conjunto difuso $\tilde{\Gamma}_A$ en $\tilde{D}$. Entonces, el soporte de A en el conjunto de transacciones difusas $\tilde{D}$ se define como

$$\text{sop}(A) = \sum_{\alpha_i \in \Lambda_{\tilde{A}}} (\alpha_i - \alpha_{i+1}) \left(\frac{|\rho_{\tilde{A}}(\alpha_i)|}{|\tilde{D}|} \right). \tag{2.10}$$

Si consideramos la parte derecha de la formula (2.10) y usamos la tabla $4ft$ asociada a los itemsets A y B para calcular el nivel de restricción α_i es fácil ver que:

$$\frac{|\rho_{\tilde{A}}(\alpha_i)|}{|\tilde{D}|} = \frac{a_i + b_i}{a_i + b_i + c_i + d_i} \tag{2.11}$$

que es justamente lo que conocemos como soporte de un itemsets cuando la base de datos es crisp. Además como hicimos notar anteriormente, $a_i + b_i + c_i + d_i$ permanece constante para cualquier nivel de restricción y es justamente el número de transacciones (en este caso difusas) de $\tilde{D}$ que notaremos por n como viene siendo habitual.

De forma parecida definiremos el soporte y la confianza para la regla de asociación difusa $A \to B$.

Definición 2.9. Sean $A, B \subseteq I$ dos itemsets disjuntos y $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}}), (\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$ las RL-representaciones asociadas a los conjuntos difusos $\tilde{\Gamma}_A$ y $\tilde{\Gamma}_B$ en $\tilde{D}$. Entonces, el soporte de la regla difusa $A \to B$ en $\tilde{D}$ se define como

$$\begin{aligned} \mathrm{Sop}(A \to B) &= \mathrm{sop}(A \wedge B) \\ &= \sum_{\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \left(\frac{|\rho_{\tilde{A} \wedge \tilde{B}}(\alpha_i)|}{|\tilde{D}|} \right) \end{aligned} \tag{2.12}$$

Definición 2.10. Sean $A, B \subseteq I$ dos itemsets disjuntos y $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}}), (\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$ las RL-representaciones asociadas a los conjuntos difusos $\tilde{\Gamma}_A$ y $\tilde{\Gamma}_B$ en $\tilde{D}$. Entonces, la confianza de la regla difusa $A \to B$ es

$$\begin{aligned} \mathrm{Conf}(A \to B) &= \sum_{\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \left(\Rightarrow_I (a_i, b_i) \right) \\ &= \sum_{\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \left(\frac{|\rho_{\tilde{A} \wedge \tilde{B}}(\alpha_i)|}{|\rho_{\tilde{A}}(\alpha_i)|} \right) \end{aligned} \tag{2.13}$$

La anterior definición tiene el inconveniente de que pueden surgir indeterminaciones de la forma "$\frac{0}{0}$" cuando $|\rho_{\tilde{A}}(\alpha_i)| = 0$. Esto ocurre cuando no existen transacciones cumpliendo el antecedente y el consecuente a la vez de entre ninguna transacción que cumple el antecedente. Por tanto, para preservar la definición 1.6 de regla de asociación difusa, en dichos casos tomaremos el valor 1 para dicha indeterminación.

En el siguiente teorema demostramos que el modelo lógico para reglas difusas generaliza al modelo para reglas de asociación crisp.

Teorema 2.1. *Sean A y B dos itemsets de una base de datos crisp D. Entonces el modelo lógico difuso presentado anteriormente coincide con el modelo para reglas de asociación crisp.*

Demostración. Para ver que la anterior formulación es una buena generalización del caso crisp, primero definiremos los conjuntos difusos asociados a A y B vistos como si fueran itemsets difusos: $\tilde{\Gamma}_A(t) = t(A), \tilde{\Gamma}_B(t) = t(B) \in [0, 1]$ donde t es una

transacción de D y $t(A), t(B) \in \{0,1\}$ vienen dados por una función indicadora de la forma:

$$t(A) = \begin{cases} 1 & \text{si } A \in t \\ 0 & \text{si } A \notin t \end{cases} \tag{2.14}$$

De forma análoga para B.

Los RL-sets asociados tanto a $\tilde{\Gamma}_A$ como a $\tilde{\Gamma}_B$ serán $\Lambda_{\tilde{A}} = \Lambda_{\tilde{B}} = \{1\}$. Y las RL-representaciones asociadas serán: $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}})$ y $(\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$ con

$$\begin{aligned} \rho_{\tilde{A}}(1) = A_1 = \{t \in D \,|\, t(A) \geq 1\} = \{t \in D \,|\, A \in t\} \\ \rho_{\tilde{B}}(1) = B_1 = \{t \in D \,|\, t(B) \geq 1\} = \{t \in D \,|\, B \in t\} \end{aligned} \tag{2.15}$$

De forma análoga podemos calcular las RL-representaciones para los conjuntos $\tilde{\Gamma}_A \wedge \tilde{\Gamma}_B$, $\tilde{\Gamma}_A \wedge \neg\tilde{\Gamma}_B$, $\neg\tilde{\Gamma}_A \wedge \tilde{\Gamma}_B$, $\neg\tilde{\Gamma}_A \wedge \neg\tilde{\Gamma}_B$. Por tanto, la tabla $4ft$ para el nivel de restricción $\alpha = 1$ coincide con la tabla $4ft$ para los itemsets crisp A y B vista en la sección 1.2:

$\mathcal{M}_1 \equiv \mathcal{M}$	B	$\neg B$
A	a_1	b_1
$\neg A$	c_1	d_1

Con respecto a las medidas para la validez de las reglas difusas, es inmediato ver que para el caso crisp coincide con los cuantificadores-4ft:

$$\sum_{\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \, (\approx (a_i, b_i, c_i, d_i)) = (1-0) \, (\approx (a_1, b_1, c_1, d_1)) \, .$$

■

Observemos que las medidas de soporte y confianza para el caso difuso coinciden con la generalización presentada en [Delgado et al., 2003a] en la que se utiliza una aproximación semántica basada en la evaluación de sentencias cuantificadas (ver también [Delgado et al., 2008d]) usando el método GD y el cuantificador relativo difuso la mayoría $Q_M(x) = x$ para evaluar las sentencias.

A continuación presentaremos dos ejemplos de los cuales el primero nos mostrará detalladamente, mediante una pequeña base de datos difusa, cuál es el proceso para calcular el soporte y la confianza de las reglas difusas usando el modelo lógico propuesto.

En la sección anterior demostramos que el factor de certeza satisface los principios deseables para una buena medida de interés. El segundo ejemplo extiende el factor de certeza para poder utilizarlo para reglas difusas y conservar así sus buenas propiedades a la hora de extraer reglas difusas.

Ejemplo 2.1. Sean $I = \{i_1, i_2, \ldots, i_6\}$ y $\tilde{D}_1$ la base de datos difusa de la Tabla 2.2. Consideremos los itemsets $A = \{i_1, i_3\}$ y $B = \{i_4\}$. Los conjuntos difusos asociados $\tilde{\Gamma}_A$ y $\tilde{\Gamma}_B$ que podemos ver en la Tabla 2.3 tendrán RL-representaciones $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}})$ y $(\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$. En la Tabla 2.4 vemos sus RL-representaciones y las RL-representaciones resultantes de aplicar la negación a ambos conjuntos difusos y la conjunción entre ellos.

	i_1	i_2	i_3	i_4	i_5
$\tilde{\tau}_1$	1	0.2	1	0.9	0.9
$\tilde{\tau}_2$	1	1	0.8	0	0
$\tilde{\tau}_3$	0.5	0.1	0.7	0.6	0
$\tilde{\tau}_4$	0.6	0	0	0.5	0.5
$\tilde{\tau}_5$	0.4	0.1	0.6	0	0
$\tilde{\tau}_6$	0	1	0	0	0

Tabla 2.2: Conjunto de transacciones difusas $\tilde{D}_1$

Calcularemos el soporte y la confianza de algunas reglas de asociación difusas en $\tilde{D}_1$ siguiendo el modelo propuesto. Si tomamos por ejemplo los itemsets $A = \{i_1, i_3\}$ y $B = \{i_4\}$, usando las anteriores RL-representaciones podemos ver cuáles son las tablas $4ft(\mathcal{M}_{\alpha_i}, \tilde{\Gamma}_A, \tilde{\Gamma}_B, \tilde{D})$ para cada nivel de restricción en $\Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}$ (Tabla 2.6). De esta forma, podemos comprobar que el soporte de la regla $A \rightarrow B$ es 0.233 usando la ecuación (2.12) y su confianza es 0.517.

	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}_3$	$\tilde{\tau}_4$	$\tilde{\tau}_5$	$\tilde{\tau}_6$
$\tilde{\Gamma}_A$	1	0.8	0.5	0	0.4	0
$\tilde{\Gamma}_B$	0.9	0	0.6	0.5	0	0

Tabla 2.3: Conjuntos difusos $\tilde{\Gamma}_A$ y $\tilde{\Gamma}_B$

α_i	$\rho_{\tilde{A}}$	$\rho_{\tilde{B}}$	$\rho_{\neg\tilde{A}}$	$\rho_{\neg\tilde{B}}$
1	ϕ	$\{\tilde{\tau}_2, \tilde{\tau}_3, \tilde{\tau}_4, \tilde{\tau}_5, \tilde{\tau}_6\}$	$\tilde{D}$	$\{\tilde{\tau}_1\}$
0.9	ϕ	$\tilde{D}$	$\tilde{D}$	ϕ
0.75	$\{\tilde{\tau}_3, \tilde{\tau}_6\}$	$\tilde{D}$	$\{\tilde{\tau}_1, \tilde{\tau}_2, \tilde{\tau}_4, \tilde{\tau}_5\}$	ϕ
0.1	$\{\tilde{\tau}_3, \tilde{\tau}_4, \tilde{\tau}_6\}$	$\tilde{D}$	$\{\tilde{\tau}_1, \tilde{\tau}_2, \tilde{\tau}_5\}$	ϕ

Tabla 2.4: *RL*-representaciones asociadas a $\tilde{\Gamma}_A$, $\tilde{\Gamma}_B$, $\neg\tilde{\Gamma}_A$, $\neg\tilde{\Gamma}_B$.

α_i	$\rho_{\tilde{A}\wedge\tilde{B}}$	$\rho_{\tilde{A}\wedge\neg\tilde{B}}$	$\rho_{\neg\tilde{A}\wedge\tilde{B}}$	$\rho_{\neg\tilde{A}\wedge\neg\tilde{B}}$
1	ϕ	ϕ	$\{\tilde{\tau}_2, \tilde{\tau}_3, \tilde{\tau}_4, \tilde{\tau}_5, \tilde{\tau}_6\}$	$\{\tilde{\tau}_1\}$
0.9	ϕ	ϕ	$\tilde{D}$	ϕ
0.75	$\{\tilde{\tau}_3, \tilde{\tau}_6\}$	ϕ	$\{\tilde{\tau}_1, \tilde{\tau}_2, \tilde{\tau}_4, \tilde{\tau}_5\}$	ϕ
0.1	$\{\tilde{\tau}_3, \tilde{\tau}_4, \tilde{\tau}_6\}$	ϕ	$\{\tilde{\tau}_1, \tilde{\tau}_2, \tilde{\tau}_5\}$	ϕ

Tabla 2.5: *RL*-representaciones asociadas a las posibles conjunciones entre los conjuntos difusos $\tilde{\Gamma}_A$, $\tilde{\Gamma}_B$, $\neg\tilde{\Gamma}_A$, $\neg\tilde{\Gamma}_B$.

Si tomamos $C = \{i_1, i_5\}$ y $E = \{i_4\}$ tendremos que el soporte de la regla difusa $C \rightarrow E$ es también 0.233, pero su confianza es igual a 1. Queremos resaltar que la confianza es 1 puesto que $\tilde{\tau}_i(C) \leq \tilde{\tau}_i(E)$ para todo $i = 1, 2, \ldots, 6$ siguiendo la definición 1.6. Si usamos la fórmula para la confianza dada en (2.13) también llegamos al mismo resultado pero teniendo en cuenta que cada vez que aparezca la indeterminación "$\frac{0}{0}$" tendremos que considerarla como 1 como ya se comentó después de la definición 2.10.

Otras reglas de asociación que pueden encontrarse en $\tilde{D}_1$ se encuentran en la siguiente tabla:

Regla	**Soporte**	**Confianza**
$\{i_1, i_2\} \rightarrow \{i_3\}$	0.2	0.8
$\{i_4\} \rightarrow \{i_5\}$	0.233	0.683

♦

Este segundo ejemplo no es más que otro ejemplo de extensión de una medida de interés crisp para reglas difusas. Extenderemos la definición del factor de certeza presentado en [Berzal et al., 2002], el cual, visto como cuantificador, lo

	a_i	b_i	c_i	d_i
$\mathcal{M}_1$	0	1	0	5
$\mathcal{M}_{0.9}$	1	0	0	5
$\mathcal{M}_{0.8}$	1	1	0	4
$\mathcal{M}_{0.6}$	1	1	1	3
$\mathcal{M}_{0.5}$	2	1	1	2
$\mathcal{M}_{0.4}$	2	2	1	1

Tabla 2.6: $4ft(\mathcal{M}_{\alpha_i}, A, B, \tilde{D})$ con $\alpha_i \in \Lambda_A \cup \Lambda_B$

denotamos por $\equiv_{FC}$ en la sección anterior después de probar teóricamente sus buenas propiedades para la extracción de reglas de asociación. Esta extensión al caso difuso será utilizada en la sección 2.3 cuando busquemos reglas difusas en un tipo particular de base de datos formado por bolsas.

Ejemplo 2.2. Sean $X, Y \subset I$ dos itemsets disjuntos en D y sea $4ft(X, Y, D) = \langle a, b, c, d\rangle$ su tabla-4ft asociada. El cuantificador $\equiv_{FC}$ asociado al factor de certeza viene dado por:

$$\equiv_{FC}(a,b,c,d) = \begin{cases} \dfrac{ad-bc}{(a+b)(b+d)} & \text{si } ad > bc \\ 0 & \text{si } ad = bc \\ \dfrac{ad-bc}{(a+b)(a+c)} & \text{si } ad < bc. \end{cases}$$

Para generalizar al caso difuso este cuantificador, es decir, si consideramos que X, Y son dos itemsets disjuntos difusos en $\tilde{D}$ sólo tenemos que considerar la tabla-4ft para cada nivel de restricción de $\Lambda_{\tilde{X}} \cup \Lambda_{\tilde{Y}}$ (la unión de los RL-sets de $\tilde{\Gamma}_X$ y $\tilde{\Gamma}_Y$) y calcular:

$$\equiv_{FC}(X \to Y) = \sum_{\alpha_i \in \Lambda_{\tilde{X}} \cup \Lambda_{\tilde{Y}}} (\alpha_i - \alpha_{i+1}) \left(\equiv_{FC}(a_i, b_i, c_i, d_i)\right) \tag{2.16}$$

donde las RL-representaciones de ambos itemsets deben estar normalizadas para que se cumpla la definición 1.6.

En particular, para la base de datos difusa $\tilde{D}_1$, y las anteriores elecciones de itemsets, tenemos los factores de certeza asociados en la Tabla 2.7.

Observemos, que en el caso de $X = \{i_1, i_5\}, Y = \{i_4\}$ tendremos que normalizar sus RL-representaciones (dividir todos los niveles de restricción por el mayor de

Regla	Soporte	Confianza	Factor de Certeza
$\{i_1, i_3\} \rightarrow \{i_4\}$	0.233	0.517	0.238
$\{i_1, i_5\} \rightarrow \{i_4\}$	0.233	1	1
$\{i_1, i_2\} \rightarrow \{i_3\}$	0.2	0.8	0.6
$\{i_4\} \rightarrow \{i_5\}$	0.233	0.683	0.59

Tabla 2.7: Factor de certeza para algunas reglas difusas en $\tilde{D}_1$

ellos, en este caso, dividiremos por 0.9) para obtener el valor 1 para el factor de certeza puesto que si $\tilde{\tau}_i(X) \leq \tilde{\tau}_i(Y)$ para todo $i = 1, \ldots, 6$ la regla de asociación es totalmente cierta. ♦

2.2.2. Otras propuestas

Definir las reglas de asociación difusas es un paso casi inmediato basándonos en la generalización de conjuntos (intervalos) por conjuntos difusos (intervalos difuso). Por lo que las reglas descubiertas en una base de datos pueden ser representadas lingüísticamente de forma más comprensiva y amigable para el usuario como vimos en la sección 1.1.2. Sin embargo, la evaluación de las reglas difusas mediante medidas de validez apropiadas no es tan sencillo [Dubois et al., 2003]. Especialmente si asumimos la semántica asociada a la regla difusa. A este respecto, se han desarrollado varias propuestas, entre las que destacamos las siguientes:

- Sudkamp establece medidas de validez para reglas difusas basándose en una clasificación de las transacciones en tres tipos: ejemplos, contraejemplos y ejemplos irrelevantes [Sudkamp, 2005]. Según el grado de pertenencia de las transacciones a estos tres conjuntos, se calculan las medidas de *confirmación* y *confianza* para reglas difusas.

- Dubois et al. [Dubois et al., 2006] proponen un método para derivar las medidas para reglas difusas de forma sistemática basándose, de nuevo, en la clasificación de los datos en ejemplos, contraejemplos y ejemplos irrelevantes asociados a una regla. En el caso difuso, la correspondiente partición será difusa y hay distintas formas de definirla dependiendo en la semántica que se quiera reflejar en la regla. Según esta propuesta, se pueden definir distintos tipos de reglas difusas, como que basan su definición en un enfoque basado en la conjunción de conjuntos o bien en un enfoque basado

en la implicación. Dichas propuestas coinciden en el caso crisp por lo que no se distinguen como ocurre con las reglas difusas (ver la versión extendida de [Dubois et al., 2006]).

- De Cock et al. proporcionan también un estudio de la semántica de las reglas difusas mediante la definición de los llamados ejemplos positivos y negativos asociados a la regla [Cock et al., 2005].

La idea clave de nuestra propuesta es, en cierto modo, la misma que en los anteriores ejemplos ya que el modelo formal nos brinda la clasificación de las tuplas en cuatro tipos distintos. La principal diferencia es que nosotros no medimos en los conjuntos en los que dividimos las tuplas, si no que primero vemos en cada nivel de restricción cuáles son las tuplas que corresponden a cada tipo y después agregamos la medida para cada nivel según la diferencia entre los niveles.

De esta forma, en cada nivel nos dará una medida de la validez de la regla que se irá agregando a lo largo de los niveles, pudiendo incluso disminuir la medida global difusa como ocurre por ejemplo en el caso de las medidas de interés cuyo rango de valores oscila en el intervalo [-1,1].

Otra propuesta interesante que también utiliza *RL*-representaciones es la dada en [Molina et al., 2009]. Los autores manejan tanto la presencia como la ausencia de items para obtener reglas de asociación crisp en cada nivel de restricción, de forma que pueden utilizar las medidas de interés usuales para reglas crisp para ver si en un determinado nivel de restricción se cumple una regla de asociación. Además proveen una medidas que sirven para resumir la información obtenida en el *RL*-set, es decir, en el conjunto de todos los niveles de restricción usando para ello la misma generalización obtenida con nuestra propuesta para el soporte, la confianza y el factor de certeza difusos. No obstante su propuesta difiere de la nuestra en que una vez que la regla crisp satisface los umbrales impuestos para dichas medidas, se considera que la regla se cumple en dicho nivel de restricción.

En la siguiente sección presentaremos varias propuestas para extraer conocimiento en base de datos donde los items pueden aparecer más de una vez en cada transacción. Este tipo de transacciones son conocidas como *bolsas*.

2.3. Extracción de Conocimiento en Bolsas

Muchas de los datos que nos podemos encontrar en la vida real involucran pares de la forma $\langle item, cantidad \rangle$. Este tipo especial de dato puede ser caracterizado usando la *teoría de bolsas o multisets*. Las reglas de asociación estudiadas hasta ahora (ver la sección 1.1) aunque surgieron en el contexto de las bolsas de la compra, no sacan toda la información de interés concerniente a la cantidad de los artículos de la bolsa, si no que más bien se han centrado en medir cuando los artículos se compraban conjuntamente en la mayoría de las transacciones. De esta forma, las reglas de asociación encontraban relaciones de la forma *la mayoría de las transacciones que contienen pan también tienen mantequilla*, y se las notaba usando una implicación $pan \rightarrow mantequilla$.

Es cierto que desde su aparición las reglas de asociación han evolucionado para desarrollar nuevos tipos de asociaciones que se ajusten a los datos que estemos manejando, como las reglas difusas, o que intentan extraer nuevos tipos de conocimiento, como las reglas graduales y las tendencias. Pero los esfuerzos por involucrar en este tipo de asociaciones nuevos factores tales como la cantidad, el tiempo, la importancia, la utilidad, etc. son aún escasos.

En esta sección ofreceremos varios procedimientos para extraer diversos tipos de conocimiento procedente de bolsas y estableceremos relaciones entre artículos involucrando su cantidad, por ejemplo reglas del tipo: *la mayoría de las bolsas que contienen una barra de pan, también contienen dos litros de leche.* La teoría de las bolsas [Yager, 1986], [Delgado et al., 2003c] pretende formalizar cómo especificar la cantidad asociada a un item en una transacción, de esta forma, una bolsa contendrá más información que una simple transacción puesto que dice cuantos items aparecen, y por tanto, una bolsa puede ser vista como una generalización del concepto de transacción. Por ello, las asociaciones que establezcamos proporcionarán una información más amplia sobre la relación entre dos conjuntos de items.

Además, este tipo especial de transacciones a las que hemos denominado bolsas pueden aparecer afectadas de incertidumbre. Más aún, es normal el uso de expresiones imprecisas que involucran cantidad, como puede ser "la mayoría de la gente que compra mucho pan también compran mucha leche". La Teoría de los conjuntos difusos [Zadeh, 1965] vuelve a ser una base excelente para expresar conocimiento vago o impreciso de forma significativa y manejable. En el caso de las reglas de asociación ya vimos cómo los conjuntos difusos ayudan en su formulación

y posterior extracción de las reglas difusas o cuantitativas difusas [Gyenesei, 2001], [Chen and Wei, 2002], [Delgado et al., 2003a]. Estas mismas herramientas pueden adaptarse para extraer información tanto en bolsas como en bolsas difusas.

En particular, la propuesta que vamos a presentar obtiene distintos tipos de relaciones entre las variables usando tanto reglas difusas [Delgado et al., 2003a] como dependencias graduales [Berzal et al., 2007], [Molina et al., 2007] analizadas con detalle en la sección 1.1.3.

El resto de la sección está organizada como sigue: primero repasaremos las definiciones de bolsa y bolsa difusa que nos harán falta para el entendimiento de nuestras propuestas. A continuación, haremos un repaso de la única propuesta que hemos encontrado que ofrece asociaciones entre bolsas. Después desarrollaremos los distintos métodos para obtener reglas difusas y dependencias graduales tanto en bolsas como en bolsas difusas. Para terminar mostraremos en un ejemplo las técnicas propuestas para aclarar los distintos procesos y los tipos de reglas obtenidas en cada uno de ellos.

2.3.1. Bolsas y Bolsas Difusas

Las llamadas *bolsas* o *multisets* fueron introducidos por R. Yager en 1986 como estructuras algebraicas parecidas a los conjuntos donde un elemento podía aparecer más de una vez [Yager, 1986]. Las bolsas son útiles para modelar situaciones como las que aparecen en bolsas de la compra aunque puedes usarse en otros contextos. Por ejemplo, en el mundo real es común encontrar que varios objetos verifiquen las mismas propiedades en un cierto contexto (espacio y tiempo), es decir, podremos decir que dichos objetos son instancias de la misma clase. Una de las aplicaciones naturales de las bolsas es la de contar cuantos objetos cumplen una determinada propiedad bajo un contexto dado [Delgado et al., 2003c].

Para definir el concepto de bolsa, en la literatura podemos encontrarnos con varias definiciones [Yager, 1986], [Delgado et al., 2003c] aunque informalmente podemos decir que una bolsa B no es más que una colección de elementos que pueden contener duplicados.

Definición 2.11. (Bolsa) Sea $I = \{i_1, \ldots, i_m\}$ un conjunto de items. Una *bolsa* B es un conjunto de items, cada uno de ellos no necesariamente distinto. Notaremos a las bolsas por un conjunto que contiene pares de la siguiente forma:

$$\{\langle b_1, q_1\rangle, \ldots, \langle b_k, q_k\rangle\} \tag{2.17}$$

donde $b_j \in I$ y q_j es un número entero no negativo para todo $1 \leq j \leq k$.

Un ejemplo de bolsa podría ser $B = \{i_1, i_1, i_3, i_3, i_3, i_7, i_7\}$ y se representaría por $B = \{\langle i_1, 2\rangle, \langle i_3, 3\rangle, \langle i_7, 2\rangle\}$.

Definición 2.12. (Bolsa difusa) Una *bolsa difusa* es una estructura algebraica cuyos elementos son ternas de la forma $\langle b_i, w_i, q_i\rangle$ donde $b_i \in I$ denota el objeto o item, $w_i \in [0, 1]$ mide la importancia, relevancia u otra característica del objeto en la bolsa, y q_i denota el número de veces que el par $\langle b_i, w_i\rangle$ aparece en la bolsa.

Las bolsas difusas son muy útiles para modelizar situaciones en las que manipulamos conjuntamente objetos y sus propiedades y además estos objetos ocurren más de una vez. Un ejemplo de uso de las bolsas difusas puede ser la representación de un conjunto de términos que constituyen la información esencial en un documento (ver el ejemplo 2.5). Así el conjunto de ternas $\langle término, importancia, frecuencia\rangle$ con $término \in documento$ es una bolsa difusa.

A un conjunto de bolsas que guardan alguna relación entre sí, le llamaremos *base de datos formada por bolsas*. En el ejemplo anterior, un documento es una base de datos formada por bolsas difusas de la forma $\langle término, importancia, frecuencia\rangle$.

Definición 2.13. (Base de datos formada por bolsas o bolsas difusas) Una *base de datos formada por bolsas (difusas)* es un conjunto $D = \{B_1, \ldots, B_n\}$ donde B_i es una bolsa (difusa) para cada $1 \leq i \leq n$.

Ejemplo 2.3. Sea I = {camisas, faldas, pantalones, calcetines, camisetas} el conjunto de los artículos de venta en una tienda de ropa. Representaremos cada artículo por a_1, a_2, a_3, a_4, a_5 respectivamente. Los siguientes conjuntos podrían ser ejemplos de bolsas:

$$B_1 = \{\langle a_1, 2\rangle, \langle a_3, 1\rangle\}$$
$$B_2 = \{\langle a_2, 3\rangle, \langle a_3, 1\rangle, \langle a_4, 2\rangle, \langle a_5, 3\rangle\}$$
$$B_3 = \{\langle a_1, 1\rangle, \langle a_2, 1\rangle, \langle a_3, 3\rangle, \langle a_4, 5\rangle\}.$$

que forman parte de una base de datos formada por bolsas a la que notaremos por D_B representada en la Tabla 2.8.

2.3.2. Anteriores Propuestas

Para nuestro conocimiento, [Hsu et al., 2004] es el único trabajo que estudia cómo obtener reglas en bases de datos formadas por bolsas. La propuesta de Hsu

D_B	camisas	faldas	pantalones	calcetines (pares)	camisetas
B_1	2	0	1	0	0
B_2	0	3	1	2	3
B_3	1	1	3	5	0
B_4	1	2	0	0	2

Tabla 2.8: Base de Datos D_1 formada por las bolsas $B_1, \ldots, B_4$.

et al. considera tres tipos diferentes de reglas que pueden ser extraídas de este tipo de transacciones, es decir, de las bolsas. Las tres clases de reglas las denominan como reglas simples, generales y semánticas y las explicamos con más detalle a continuación.

1. Las reglas *simples* son reglas de la forma "(zumo = 3)→(leche fresca = 2)", cuyo significado es que si compramos 3 litros de zumo, entonces es probable que compremos 2 litros de leche fresca. En estas reglas los operadores que aparecen en ambos lados de la regla deben ser iguales, pudiendo ser alguno de los del conjunto $\{=, \geq\}$.

2. Las reglas *generales* son del tipo "(grabador de video = 1)→(cinta de video $\geq$ 6)", representando que si compramos un grabador de video, entonces seguramente compremos al menos 6 cintas de video. Los operadores utilizados en este caso pueden ser diferentes siendo alguno de estos tres tipos $\{=, \geq, \leq\}$.

3. Las reglas *semánticas* usan términos semánticos para describir la cantidad. Una regla semántica podría ser del tipo "(mucha cantidad de leche fresca)→(poca cantidad de coca-cola)". Las reglas semánticas ayudan a ver si las relaciones entre los productos son complementarias o bien competitivas. En [Hsu et al., 2004] presentan dos propuestas para definir los términos semánticos. La primera de ellas usa intervalos predefinidos por el usuario, y en el segundo define términos semánticos mediante conjuntos difusos.

Además, los autores ofrecen un algoritmo para obtener cada tipo de regla, probándolos para datos sintéticos y obteniendo mejores resultados cuando usan las reglas simples y las semánticas con términos difusos.

El uso de las reglas simples y generales tienen los mismos inconvenientes que las reglas con atributos cuantitativos debido al gran número de items que se

generan y por consiguiente también el alto número de reglas extraídas. Cuando se usan intervalos en las reglas generales también surge el problema de la frontera [Kuok et al., 1998] como explicamos en la sección 1.1.2. Sin embargo la propuesta de las reglas semánticas usando términos difusos es más natural y corresponde adecuadamente con nuestra percepción de la realidad, ya que podemos distinguir varios rangos de cantidad ofreciendo una semántica beneficiosa para el usuario.

Las propuestas que vamos a desarrollar [Delgado et al., 2008c] van en esta última línea de trabajo, pero en vez de usar el método propuesto por Hsu et al. que se centraba en analizar bolsas de la compra, presentaremos un nuevo punto de vista más general para obtener reglas de asociación en bases de datos formadas por bolsas que pueden proceder de cualquier ámbito, y que, como ya hemos visto, ofrecen una generalización de las bases de datos transaccionales. Como aplicación a destacar veremos su utilización en minería de textos.

También existen otros trabajos que utilizan datos cuantitativos que podrían ser generalizados mediante bolsas para extraer reglas secuenciales difusas (ver por ejemplo [Chen and Huang, 2006], [Hong et al., 2006]). Las bolsas difusas se han aplicado en consultas flexibles en [Rocacher, 2003]. Estos trabajos están un poco lejos de nuestro objetivo principal que es la extracción de reglas difusas y dependencias graduales tanto en bolsas como en bolsas difusas.

2.3.3. Minería de Datos en Bolsas

El método que vamos a presentar se basa principalmente en dos ideas básicas:

- La primera es transformar de forma adecuada una base de datos formada por bolsas en una base de datos formada por transacciones difusas. Para ello el usuario deberá definir una serie de conjuntos difusos para medir de una forma significativa la cantidad del usuario. También se podrían definir los conjuntos difusos teniendo en cuenta la distribución de las cantidades de cada ítem.

- La segunda es utilizar las herramientas que están a nuestro alcance para obtener distintos tipos de información de la base de datos transaccional difusa. Nosotros distinguiremos esencialmente el uso de reglas difusas y de dependencias graduales.

Para realizar una transformación adecuada deberemos tener en cuenta tanto la información proporcionada por el usuario, como la naturaleza de los datos

y también el tipo de conocimiento que vayamos a extraer. A continuación detallaremos las distintas propuestas.

Transformación de Bolsas a Transacciones Difusas

En este apartado presentaremos un método para transformar cualquier base de datos formada por bolsas o bolsas difusas en una base de datos transaccional difusa (ver [Delgado et al., 2008c]), que después utilizaremos para extraer reglas difusas como vimos en la sección 1.1.2.

Sean $D = \{B_1, \ldots, B_n\}$ una base de datos formada por bolsas e $I = \{i_1, \ldots, i_m\}$ un conjunto de items. Transformamos D en una base de datos transaccional difusa que notaremos por $\tilde{D}$ donde los nuevos items son pares de la forma $< i, L >$ con $i \in I$ y L una etiqueta lingüística que represente una cantidad asociada al ítem i. Por tanto, la base de datos difusa $\tilde{D}$ tiene como conjunto de items asociado: $\{< i_1, L_1 >, < i_1, L_2 >, \ldots, < i_m, L_p >\}$ donde i representa un ítem de I y $L \in \{L_{1i}, L_{2i}, \ldots, L_{si}\}$.

De esta forma por cada ítem i que hubiera inicialmente aparecerán tantos items nuevos como etiquetas asociadas a cada item digamos s_i. El número de etiquetas definidas por ítem no tiene que ser el mismo y las etiquetas generalmente serán distintas para cada ítem.

El valor de estos nuevos items de la forma $< i, L >$ en la base de datos creada $\tilde{D}$ es un valor real en el intervalo [0, 1] puesto que $\tilde{D}$ está formado por transacciones difusas. Cada transacción resultante proviene de una bolsa donde para cada ítem $< i, L >$ cuyo valor en la transacción es el valor correspondiente de la evaluación de la cantidad en el conjunto difuso dado por la etiqueta L. Formalmente si en la bolsa B_j el ítem i tenía asociada la cantidad q_{ij} entonces, en la transacción $t_j \in \tilde{D}$ el ítem $< i, L >$ tendrá el valor $\mu_L(q_{ij})$.

Huelga decir que la base de datos difusa obtenida en cada caso dependerá de los conjuntos difusos que definamos para cada ítem.

En el ejemplo 2.4 que presentaremos en la siguientes sección, veremos con más claridad cuál es el proceso para convertir bolsas en transacciones difusas.

Reglas Difusas en Bolsas

Para obtener reglas difusas en un conjunto de bolsas D, primero deberemos definir los conjuntos difusos asociados a las etiquetas lingüísticas que denotan cantidad para cada ítem que aparece en D. Después procederemos a transformar

D_2	mantequilla (kg)	pan (un.)	leche (l)	galletas (kg)
B_1	1	1	13	1.75
B_2	1.25	2	18	2
B_3	0.25	1	5	1
B_4	0.25	5	1	0.5
B_5	0.1	2	0	0.5
B_6	0.5	0	5	1

Tabla 2.9: Base de datos D_2 formada por bolsas.

D en una base de datos transaccional difusa $\tilde{D}$ y por último procederemos a extraer las reglas difusas en esta última.

Para ver más claro todo el proceso, lo haremos paso por paso en el siguiente ejemplo.

Ejemplo 2.4. Sea D_2 una base de datos formada por seis bolsas mostrada en la Tabla 2.9. Para cada ítem de $I_2 = \{\text{mantequilla, pan, leche, galletas}\}$ definimos un conjunto de términos lingüísticos referentes a sus cantidades usando conjuntos difusos como podemos observar en la Figura 2.4.

Usando estos conjuntos difusos transformamos D_2 en el conjunto de transacciones difusas dado en la Tabla 2.10 y que notaremos por $\tilde{D}_2$.

En la Tabla 2.10 hemos cambiado filas por columnas por falta de espacio, y hemos puesto un identificado a cada nuevo ítem generado. Observemos que para cada ítem inicial el conjunto de etiquetas asociadas es distinto puesto que cada uno tiene asociado una forma diferente de medir la cantidad. El número de etiquetas dependerá de la *granularidad* que el usuario desee para hacer el análisis así como de su experiencia particular. Una granularidad alta se corresponde con un número bajo de etiquetas. El dominio está poco particionado y tiene el inconveniente de que se pueda perder algo de expresividad. Por el contrario, una granularidad baja se asocia con un número alto de etiquetas. Una granularidad demasiado baja puede provocar un aumento de la complejidad en la descripción del dominio.

En este ejemplo, hemos considerado tres etiquetas para cada ítem aunque en principio no existe ninguna restricción para que el número de etiquetas coincida.

El siguiente paso consiste en la extracción de las reglas difusas. Para ello calculamos el soporte de todos los items y después buscamos el conjunto de itemsets

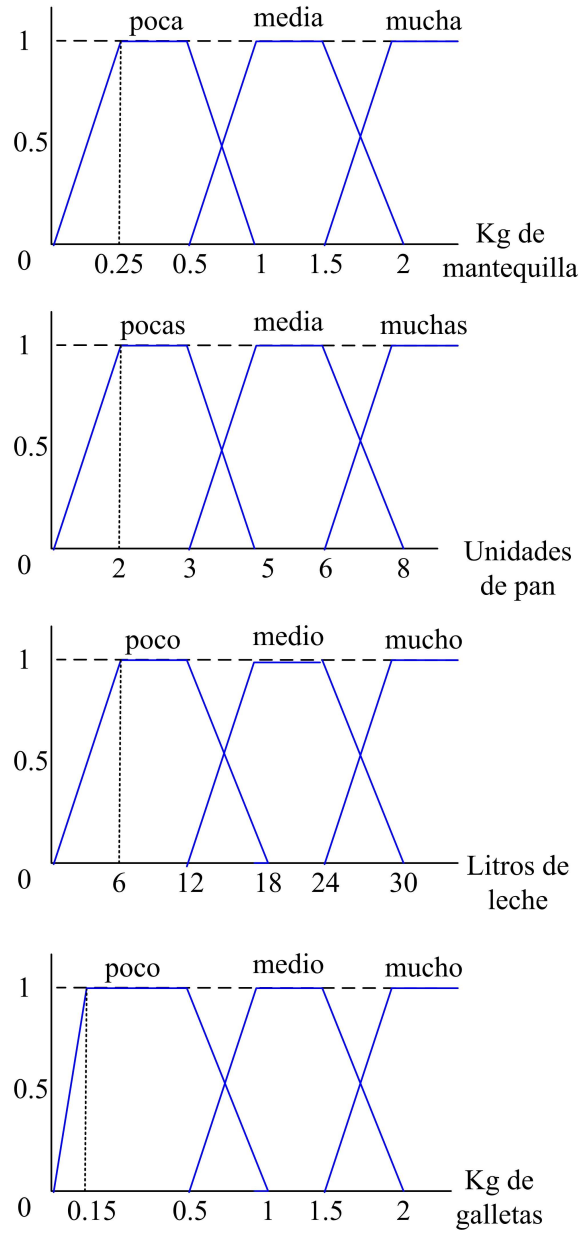

Figura 2.4: Etiquetas lingüísticas para los items de I_2.

Id	<Item, término lingüístico>	$\tilde{\tau}_1$	$\tilde{\tau}_2$	$\tilde{\tau}_3$	$\tilde{\tau}_4$	$\tilde{\tau}_5$	$\tilde{\tau}_6$
i_1	<mantequilla, poca>	0	0	1	1	0.4	1
i_2	<mantequilla, media>	1	1	0	0	0	0
i_3	<mantequilla, mucha>	0	0	0	0	0	0
i_4	<pan, poco>	0.5	1	0.5	0	1	0
i_5	<pan, medio>	0	0	0	1	0	0
i_6	<pan, mucho>	0	0	0	0	0	0
i_7	<leche, poca>	0.833	0	0.833	0.167	0	0.833
i_8	<leche, media>	0.167	1	0	0	0	0
i_9	<leche, mucha>	0	0	0	0	0	0
i_{10}	<galletas, pocas>	0	0	0	1	1	0
i_{11}	<galletas, media>	0.5	0	1	0	0	1
i_{12}	<galletas, muchas>	0.5	1	0	0	0	0

Tabla 2.10: Conjunto $\tilde{D}_2$ de transacciones difusas asociadas a D_2.

frecuentes pero con la condición de que en cada itemsets el ítem inicial $i \in I$ sólo puede ser incluido una vez. En este ejemplo, el ítem i_1 no podrá pertenecer al mismo itemsets en el que ya esté i_2 e i_3. La Tabla 2.11 muestra el soporte de todos los items de $\tilde{D}_2$.

Items	$\{i_3\}, \{i_6\}, \{i_9\}$	$\{i_5\}$	$\{i_8\}$	$\{i_{12}\}$
Soporte	0	0.167	0.195	0.25

Items	$\{i_2\}, \{i_{10}\}$	$\{i_{11}\}$	$\{i_7\}$	$\{i_4\}$	$\{i_1\}$
Soporte	0.333	0.417	0.444	0.5	0.567

Tabla 2.11: Soporte asociado a los items de $\tilde{D}_2$.

Teniendo en cuenta la anterior restricción, podemos calcular el soporte difuso asociado a los posibles k-itemsets, donde k indica el número de items en el itemsets. En la Tabla 2.12 mostramos los 2-itemsets y 3-itemsets que superan el umbral para el soporte (difuso) fijado como $minsupp = 0.2$. Recordemos que el soporte difuso de un itemsets es el considerado por la extensión difusa del caso crisp, en la ecuación (2.10) de la sección 2.2.

2,3-Itemsets	Soporte
$\{i_1, i_7\}$	0.306
$\{i_1, i_{10}\}$	0.233
$\{i_1, i_{11}\}$	0.333
$\{i_2, i_4\}$	0.25
$\{i_2, i_{12}\}$	0.25
$\{i_4, i_{12}\}$	0.25
$\{i_7, i_{11}\}$	0.361
$\{i_1, i_7, i_{11}\}$	0.278
$\{i_2, i_4, i_{12}\}$	0.25

Tabla 2.12: Soporte difuso de los 2,3-itemsets cuyo soporte ≥ 0.2 .

Para obtener las reglas de asociación usaremos las medidas del soporte y el factor de certeza que han sido debidamente generalizadas para el caso difuso usando el modelo formal y la representación mediante niveles de restricción

Regla Difusa	**Soporte**	**Factor de Certeza**
$\{i_1\} \rightarrow \{i_7\}$	0.306	0.155
$\{i_7\} \rightarrow \{i_1\}$	0.306	0.268
$\{i_1\} \rightarrow \{i_{10}\}$	0.233	0.1
$\{i_{10}\} \rightarrow \{i_1\}$	0.233	0.307
$\{i_1\} \rightarrow \{i_{11}\}$	0.333	0.314
$\{i_{11}\} \rightarrow \{i_1\}$	0.333	0.614
$\{i_2\} \rightarrow \{i_4\}$	0.25	0.5
$\{i_4\} \rightarrow \{i_2\}$	0.25	0.25
$\{i_2\} \rightarrow \{i_{12}\}$	0.25	0.667
$\{i_{12}\} \rightarrow \{i_2\}$	0.25	1
$\{i_4\} \rightarrow \{i_{12}\}$	0.25	0.333
$\{i_{12}\} \rightarrow \{i_4\}$	0.25	1
$\{i_7\} \rightarrow \{i_{11}\}$	0.361	0.686
$\{i_{11}\} \rightarrow \{i_7\}$	0.361	0.7
$\{i_2, i_4\} \rightarrow \{i_{12}\}$	0.25	1
$\{i_1, i_7\} \rightarrow \{i_{11}\}$	0.278	0.885

Tabla 2.13: Reglas difusas en $\tilde{D}_2$.

(consultar las ecuaciones (2.12) y (2.16)). La Tabla 2.13 contiene algunas de las reglas que pueden obtenerse en $\tilde{D}_2$ junto con sus valores de soporte y factor de certeza difusos asociados.

Queremos resaltar que las reglas extraídas tienen un significado interesante y pueden ser de gran utilidad para el usuario. Por ejemplo, la regla difusa $\{i_7\} \rightarrow \{i_1\}$ nos dice que "la mayoría de las bolsas que contienen poca cantidad de leche también contienen poca cantidad de mantequilla". Con este tipo de reglas no sólo obtenemos que los items 'leche' y 'mantequilla' tienden a ocurrir conjuntamente, si no que también obtenemos una relación entre las cantidades de cada producto.

El anterior ejemplo es el clásico ejemplo donde podemos encontrar bolsas y por consiguiente podemos aplicar el método que se ha detallado. Pero las bolsas son también adecuadas para modelizar otro tipo de situaciones. El siguiente ejemplo da muestra de ello mediante el uso de bolsas difusas en minería de textos.

Ejemplo 2.5. Algunas de las técnicas de minería de datos han sido adaptadas para su uso en minería de textos como podemos apreciar en [Delgado et al., 2002a], [Delgado et al., 2002b]. En estos trabajos los autores utilizan reglas difusas para particularizar o generalizar consultas en colecciones de documentos. En este ejemplo, vamos a centrarnos en este contexto.

Aquí, las transacciones serán los documentos y los items, los términos que aparezcan en ellos. Estos items tienen asociado un grado de importancia o relevancia difuso y una frecuencia de aparición en cada documento. Por tanto, cada documento d_j podemos representarlo por un conjunto de ternas de la forma:

$$d_j = \{\langle i_{jk}, w_{jk}, f_{jk}\rangle, k = 1, \ldots, p\} \tag{2.18}$$

donde $i_{jk} \in I$ es el k-ésimo ítem en el j-ésimo documento, $w_{jk} \in [0, 1]$ es el grado de importancia asociado al ítem i_{jk} y f_{jk} es el número de apariciones del ítem en el documento d_j.

Para medir la relevancia de los términos en un documento existen varias propuestas que usan algunas características de interés en el documento. Las más usadas, identificadas por primera vez por Luhn y Edmundson, son las características: temática, lugar, cabecera, especial [Luhn, 1958], [Edmundson, 1969]. Mediante una suma normalizada de dichas características (ver por ejemplo [Ruiz and Bailón, 2008]) conseguimos una medida de la importancia de un término en un documento.

Observemos que la terna considerada en (2.18) es justamente una bolsa difusa, y una colección de documentos podemos verlos, por tanto, como una base de datos formada por bolsas difusas. Como procedimos en el anterior ejemplo, para extraer las reglas de asociación difusas primero transformaremos esta colección en un conjunto de transacciones difusas. En este caso, como tenemos bolsas difusas podemos proceder de distintas formas según el tipo de información que queramos obtener:

(i) La forma más sencilla sería procediendo como en el anterior ejemplo. Es decir, consideramos como items iniciales los pares $< i, w >$ y, como cantidades, sus respectivas frecuencias. De este modo definiríamos etiquetas lingüísticas a las frecuencias del tipo: no aparece, aparece poco, aparece bastante, aparece muy frecuentemente, y el grado de cumplimiento del nuevo ítem en cada transacción difusa sería $\mu_{L_f}(f)$ donde f es la frecuencia y L_f es la etiqueta lingüística.

(ii) Otra posibilidad podría ser en el caso anterior tomar como nuevo ítem el par $\langle i, L_f \rangle$ siendo L_f la etiqueta que mejor se correspondiese con la frecuencia del item y el valor difuso asociado sería el grado de importancia dado por w_i.

(iii) Una posibilidad aunque un poco más compleja que las anteriores, sería llevar a cabo dos tipos de *agrupamiento*. Uno para las frecuencias como en el caso (i) y otro para el grado de importancia. El primer agrupamiento se haría con conjuntos difusos que denotan frecuencia y el segundo con conjuntos difusos que denoten calidad o importancia. Con esta doble agrupación los nuevos items de la base de datos transaccional difusa serían de la forma $\langle i, L_w, L_f \rangle$, donde L_w, L_f son etiquetas lingüísticas para el grado y para la frecuencia respectivamente. El valor que dicho ítem tendría en una transacción se calcularía usando una t-norma entre los grados de pertenencia asociados a los valores w, f en su correspondiente etiqueta. Este valor nos daría el grado de cumplimiento de las dos características (importancia y frecuencia).

Una vez transformada la colección de documentos en un conjunto de transacciones difusas, procederemos a la extracción de las reglas difusas [Delgado et al., 2002a] que nos ayudarán en el proceso de particularización o generalización de las consultas realizadas en dichos documentos.

Dependencias Graduales en Bolsas

Las llamadas *dependencias graduales* (introducidas con más detalle en la sección 1.1.3) miden una variación en la relación existente entre dos tipos de objetos [Hüllermeier, 2002]. Este tipo de dependencias permiten expresar una "tendencia" en los datos . Un ejemplo de dependencia gradual sería la regla *cuanto mayor el salario, mejor el puesto de trabajo*, significando que cuando el salario de una persona crece, esta posee un mayor rango en su trabajo. Los cambios en el grado de pertenencia en las dependencias graduales pueden ser de dos tipos: *mayor* o *menor*. Si X, Y son dos atributos y A, B son conjuntos difusos definidos en X e Y respectivamente, podremos considerar cuatro tipos diferentes de dependencias graduales: *cuanto mayor* X *es* A*, mayor* Y *es* B (expresado como $(>, X, A) \rightarrow (>, Y, B)$), *cuanto mayor* X *es* A*, menor* Y *es* B expresado como $(>, X, A) \rightarrow (<, Y, B)$), etc.

Formalmente una dependencia gradual [Berzal et al., 2007] es una regla de la forma $(\rhd_1, X, A) \rightarrow (\rhd_2, Y, B)$, con $\rhd_1, \rhd_2 \in \{<, >\}$. La dependencia se cumplirá en una base de datos D si y sólo si para cualesquiera $(x, y), (x', y') \in D$ con $x, x' \in X$ e $y, y' \in Y$, el cumplimiento de $A(x) \rhd_1 A(x')$ implica que $B(y) \rhd_2 B(y')$. Para extraer este tipo de dependencias usando reglas de asociación, tendremos que hacerlo en un conjunto apropiado de transacciones de D.

Nuestro propósito es buscar el mejor método para encontrar dependencias graduales en el caso de tener una base de datos formada por bolsas que aproveche de la mejor manera posible toda la información almacenada. Siguiendo las dos ideas básicas que sigue nuestra propuesta, ofreceremos cuatro métodos distintos [Delgado et al., 2008c] para transformar el conjunto de bolsas en una base de datos transaccional para después extraer distintos tipos de dependencias graduales que nos mostrarán diversas formas de extraer información interesante y útil para el usuario. De nuevo, para una mejor comprensión se explicará el procedimiento en el conjunto de bolsas D_2 que tomamos como ejemplo.

Dependencias Graduales Crisp en Bolsas

En el ambiente fijado anteriormente, cuando queremos extraer dependencias graduales crisp, el conjunto apropiado de transacciones es el que definimos como GI^D formado por items de la forma $[>, X, A], [<, X, A], [>, Y, B], [<, Y, B]$ que expresan las dos posibles tendencias de los atributos X e Y con respecto a las restricciones marcadas por los conjuntos difusos A y B respectivamente. También definimos el conjunto de transacciones, que notaremos por GT^D, como aquel que contiene los items de GI^D obtenidos de D como sigue: para cada par

$o = (x, y)$, $o' = (x', y') \in D$ existe una transacción $gt_{oo'} \in GT^D$ tal que $[\rhd, X, A]$ pertenecerá a $gt_{oo'}$ si, y sólo si, $A(x) \rhd A(x')$ con $\rhd \in \{<, >\}$ (de forma similar para $[\rhd, Y, B]$). Entonces, diremos que la dependencia gradual $(\rhd_1, X, A) \rightarrow (\rhd_2, Y, B)$ se cumple en D si, y sólo si la regla de asociación (crisp) $[\rhd_1, X, A] \rightarrow [\rhd_2, Y, B]$ se cumple en GT^D. Las medidas de soporte y confianza definidas para la extracción de reglas de asociación podrán ser utilizadas para validar la anterior dependencia gradual.

Basándonos en esta propuesta, propondremos dos métodos distintos para extraer dependencias graduales crisp en un conjunto de bolsas.

Método 1.

El primer método utilizará directamente la información contenida en las bolsas para extraer conocimiento sobre la tendencia a aumentar o disminuir las cantidades almacenadas en las bolsas. De nuevo utilizaremos la base de datos formada por bolsas D_2 para explicarlo.

Sea D_2 un conjunto de bolsas (ver Tabla 2.9). El primer paso consiste en definir un conjunto de items al que notaremos por GI^1 de la forma

$$[>, b_1], [<, b_1], \ldots, [>, b_m], [<, b_m]$$

que expresan las dos posibles tendencias de cada ítem de la bolsa, es decir, $b_1 =$ mantequilla, $b_2 =$ pan, $b_3 =$ leche y $b_4 =$ galletas. El conjunto de transacciones GT^1 se obtiene de D_2 como sigue: para cada par de bolsas (B_j, B_k) con $j \neq k$ definimos una nueva transacción a la que notaremos por gt_{jk} donde un ítem $[\rhd, b_i] \in GI^1$ se cumplirá en gt_{jk} si y sólo si $q_i(B_j) \rhd q_i(B_k)$ donde $\rhd \in \{<, >\}$ y $q_i(B_j)$ denota la cantidad asociada al ítem b_i en la bolsa B_j.

El segundo paso consiste en la extracción de las dependencias graduales (crisp) en GT^1. Para ello podrían usarse las medidas de soporte y confianza, pero nosotros utilizaremos el soporte y el factor de certeza como ya hicimos en las anteriores secciones.

Recordemos que el soporte de un ítem $[<, b_i]$ es el mismo que el de $[>, b_i]$ (ver [Molina et al., 2007]), por tanto sólo tendremos que considerar dos tipos de reglas en vez de las cuatro posibles, es decir, nos centraremos en la búsqueda de las reglas del tipo: $[>, b_i] \rightarrow [>, b_j]$ y $[>, b_i] \rightarrow [<, b_j]$. La Tabla 2.14 muestra algunas de las dependencias graduales que podemos extraer de D_2 siguiendo este método.

Método 2.

Reglas	Soporte	Factor de Certeza
$(>, mantequilla) \rightarrow (>, leche)$	0.433	0.866
$(>, galletas) \rightarrow (>, mantequilla)$	0.4	0.856
$(>, galletas) \rightarrow (>, leche)$	0.433	1

Tabla 2.14: Algunas Dependencias Graduales Crisp obtenidas en D_2.

El segundo método obtendrá las dependencias graduales mediante una doble transformación del conjunto de bolsas. La idea es muy simple, primero el conjunto de bolsas se convertirá en un conjunto de transacciones difusas y después la base de datos difusa se transformará en un conjunto de transacciones del tipo GT como hemos visto en el anterior método. Este método nos será útil cuando queramos obtener dependencias graduales con una mayor granularidad. Con un ejemplo lo veremos más claro.

Sea D_2 la base de datos formada por bolsas dada por la Tabla 2.9. El primer paso consistirá en transformar D_2 en un conjunto de transacciones difusas como ya hicimos en la sección anterior, quedando como resultado la base de datos difusa $\tilde{D}_2$ de la Tabla 2.10. El segundo paso será definir el conjunto de items que notaremos por GI^2 que expresarán los dos tipos de tendencia de cada item i_k de $\tilde{D}_2$, quedando

$$GI^2 = [>, i_1], [<, i_1], \ldots, [>, i_{12}], [<, i_{12}]$$

donde $i_1 = \langle mantequilla, poca \rangle$, $i_2 = \langle mantequilla, media \rangle$, etc. El conjunto de transacciones (crisp) GT^2 se obtendrá de $\tilde{D}_2$ como sigue: para cada par de transacciones $(\tilde{\tau}_j, \tilde{\tau}_k)$ con $j \neq k$ obtenemos una nueva transacción que notaremos por gt_{jk} donde un ítem $[\rhd, i] \in GI^2$ se cumplirá en dicha transacción si, y sólo si, $\tilde{\tau}_j(i) \rhd \tilde{\tau}_k(i)$ donde $\rhd \in \{<, >\}$ y $\tilde{\tau}_j(i) \in [0, 1]$ denota el valor asociado al ítem i en la transacción $\tilde{\tau}_j$.

Una vez efectuada dicha transformación, el último paso es la extracción de las dependencias graduales, que de nuevo lo haremos buscando las reglas de asociación en GT^2 usando el soporte y el factor de certeza como medidas de interés.

Al igual que ocurría en el Ejemplo 2.4 deberemos tener en cuenta que una regla no podrá contener más de una vez el ítem inicial $b_i \in I$. En este ejemplo, i_1 no podría pertenecer a una regla en la que i_2 o i_3 estuviesen involucrados. La Tabla 2.15 muestra algunas de las dependencias graduales que pueden obtenerse en D_2 usando este método.

Reglas	Soporte	FC
$(>, mantequilla, media) \rightarrow (>, leche, media)$	0.267	1
$(>, mantequilla, media) \rightarrow (>, galletas, muchas)$	0.267	1
$(>, leche, poca) \rightarrow (>, galletas, media)$	0.3	0.713
$(>, leche, media) \rightarrow (>, galletas, muchas)$	0.3	1
$(<, leche, poca) \rightarrow (>, pan, small)$	0.267	0.545
$(>, mantequilla, poca) \rightarrow (<, pan, poco)$	0.267	0.545
$(<, leche, media) \rightarrow (>, mantequilla, poca)$	0.267	0.825
$(<, galletas, muchas) \rightarrow (>, mantequilla, poca)$	0.267	0.825

Tabla 2.15: Dependencias Graduales en D_2 usando el Método 2.

Dependencias Graduales Difusas en Bolsas

En la Sección 1.1.3 vimos que desde datos totalmente distintos podíamos obtener el mismo conjunto de dependencias graduales crisp. Por este motivo se hace indispensable imponer una forma de medir el grado de cumplimiento de la tendencia de los items, surgiendo como consecuencia las dependencias graduales difusas. Recordemos el mecanismo de extracción de este tipo de reglas.

Sea $GI^D = \{[>, X, A], [<, X, A], [>, Y, B], [<, Y, B]\}$ un conjunto de items donde X, Y son atributos y A, B son conjuntos difusos definidos en X e Y respectivamente. Sea GT^D un conjunto de transacciones difusas que contienen los items de GI^D. GT^D se obtendrá de D como sigue: para cualesquiera $o = (x, y)$, $o' = (x', y') \in D$ se define una transacción difusa $gt_{oo'} \in GT^D$ tal que $gt_{oo'}$ es la función dada en la ecuación (2.19), con $\rhd \in \{<, >\}$.

$$gt_{oo'}([\rhd, X, A]) = \begin{cases} |A(x) - A(x')| & \text{si } A(x) \rhd A(x') \\ 0 & \text{en otro caso.} \end{cases} \tag{2.19}$$

Para cada regla difusa $[\rhd_1, X, A] \rightarrow [\rhd_2, Y, B]$ encontrada en GT^D, se define una *dependencia gradual difusa* $(\rhd_1, X, A) \rightarrow (\rhd_2, Y, B)$ en D, i.e., las reglas difusas obtenidas en GT^D definen un tipo particular de dependencia gradual difusa en D.

Para medir el cambio del grado de cumplimiento de las dependencias graduales en un conjunto de bolsas proponemos a continuación dos métodos distintos.

Método 3.

Este tercer método se basa en la idea de transformar una base de datos formada por bolsas en un base de datos transaccional difusa usando términos lingüísticos para representar las cantidades asociadas a un ítem, y después aplicamos el método presentado en la sección 1.1.3 para obtener dependencias graduales difusas midiendo el grado de variación de cada ítem usando la función vista en la ecuación (1.12).

El primer paso consiste en transformar la base de datos formada por bolsas D_2 en la base de datos difusa $\tilde{D}_3$ formada por las transacciones difusas del mismo modo que hicimos en el método 2 visto anteriormente.

El segundo paso sigue el siguiente procedimiento: transformamos los items de la forma $\langle b_i, L\rangle$ en $[<, b_i, L]$ y $[>, b_i, L]$, y consideramos las transacciones $gt_{jk} \in GT^3$ donde

$$gt_{jk}([\rhd, b_i, L]) = \begin{cases} |\tilde{\tau}_j(\langle b_i, L\rangle) - \tilde{\tau}_k(\langle b_i, L\rangle)| & \text{si } \tilde{\tau}_j(\langle b_i, L\rangle) \rhd \tilde{\tau}_k(\langle b_i, L\rangle) \\ 0 & \text{en otro caso} \end{cases} \tag{2.20}$$

donde $\rhd \in \{<, >\}$.

Reglas	**Soporte**	**FC**
$(>, leche, media) \rightarrow (>, galletas, muchas)$	0.172	0.914
$(>, pan, medio) \rightarrow (>, galletas, pocas)$	0.133	0.727
$(>, pan, small) \rightarrow (<, galletas, media)$	0.15	0.708
$(>, mantequilla, media) \rightarrow (>, galletas, muchas)$	0.2	0.681
$(>, galletas, media) \rightarrow (>, leche, poca)$	0.205	0.524

Tabla 2.16: Dependencias graduales difusas en D_2 usando el Método 3.

El último paso es evaluar las medidas de soporte y factor de certeza para las dependencias graduales encontradas en el nuevo conjunto de transacciones difusas GT^3 usando las definiciones difusas dadas en la sección 2.2. De nuevo, deberemos tener en cuenta que en una regla no puede haber más de un ítem $i \in I$ como hicimos en los casos anteriores. La Tabla 2.16 muestra algunas dependencias graduales difusas que pueden extraerse en el conjunto de bolsas D_2.

Método 4

Este último método difiere bastante de los otros tres presentados hasta ahora. La principal diferencia es que no obtendremos el conjunto de transacciones difusas como paso intermedio para obtener las dependencias graduales. La idea básica es medir en primer lugar el cambio de variación directamente desde las cantidades que tenemos asociadas a cada ítem en las bolsas y después *fuzzificamos* los resultados obtenidos. Para aclarar el proceso, reproduciremos cada paso en el conjunto de bolsas D_2.

En primer lugar, transformaremos los items b_i en $[<, b_i]$ y $[>, b_i]$, y después definimos el conjunto de transacciones $gt_{jk} \in GT^4$ como

$$gt_{jk}([\rhd, b_i]) = \begin{cases} |q_i(B_j) - q_i(B_k)| & \text{si } q_i(B_j) \rhd q_i(B_k) \\ 0 & \text{en otro caso} \end{cases} \tag{2.21}$$

donde $\rhd \in \{<, >\}$ y $q_i(B_j)$ es la cantidad asociada al ítem b_i en la bolsa B_j. Para obtener una transacción difusa, en este método, usaremos un conjunto difuso (diferente para cada ítem) que nos mida el grado de variación de la cantidad del ítem.

La Tabla 2.17 contiene una parte de las transacciones de GT^4, en particular aquellas transacciones que involucran a los items $[>, mantequilla]$ y $[>, leche]$. Una vez obtenida GT^4 usando los conjuntos difusos definidos en la Figura 2.5 podemos definir el conjunto de transacciones difusas $\widetilde{GT}^4$, donde los items coinciden con los de GT^4 pero su valor en cada transacción es la evaluación de sus valores en GT^4 en sus correspondientes conjuntos difusos, i.e. $\widetilde{GT}^4$ tendrá como items a $[\rhd, b_i]$, y sus transacciones $\widetilde{gt}_{jk}$ serán de la forma

$$\widetilde{gt}_{jk} = \mu_L(gt_{jk}([\rhd, b_i])) \tag{2.22}$$

donde L es el conjunto difuso asociado a b_i.

La Tabla 2.18 contiene las transacciones de $\widetilde{GT}^4$ que corresponden a la transformación mediante los conjuntos difusos de la Figura 2.5 del trozo de la Tabla 2.17.

Por último extraemos las reglas de asociación difusas en el conjunto de transacciones difusas $\widetilde{GT}^4$ usando las medidas de soporte y factor de certeza visto en la sección 2.2. Algunas de las dependencias graduales difusas que podemos obtener usando este método en D_2 se encuentran en la Tabla 2.19.

	$[>, mantequilla]$	$[>, leche]$
gt_{12}	0	0
gt_{13}	0.75	8
gt_{14}	0.75	12
gt_{15}	0.9	13
gt_{16}	0.5	8

Tabla 2.17: Parte de GT^4

	$[>, mantequilla]$	$[>, leche]$
$\tilde{\Gamma}_{12}$	0	0
$\tilde{\Gamma}_{13}$	0.75	0.533
$\tilde{\Gamma}_{14}$	0.75	0.8
$\tilde{\Gamma}_{15}$	0.9	0.867
$\tilde{\Gamma}_{16}$	0.5	0.533

Tabla 2.18: Parte de $\widetilde{GT}^4$.

Discusión

En este apartado analizaremos con más profundidad las distintas propuestas para extraer dependencias graduales en bases de datos formadas por bolsas.

- El primer método se basa en la idea presentada en [Berzal et al., 2007] para extraer dependencias graduales (crisp), obteniendo reglas de la forma:

  > *"cuanto mayor (menor) es la cantidad de pan, mayor (menor) es la cantidad de mantequilla".*

  Este tipo de dependencias sólo especifica cómo es la tendencia existente en los datos.

- El método 2 también extrae dependencias graduales crisp, pero utiliza una fuzzificación de los datos en vez del conjunto de bolsas inicial. Esta propuesta nos da dependencias con cierta granularidad, es decir, reglas del tipo:

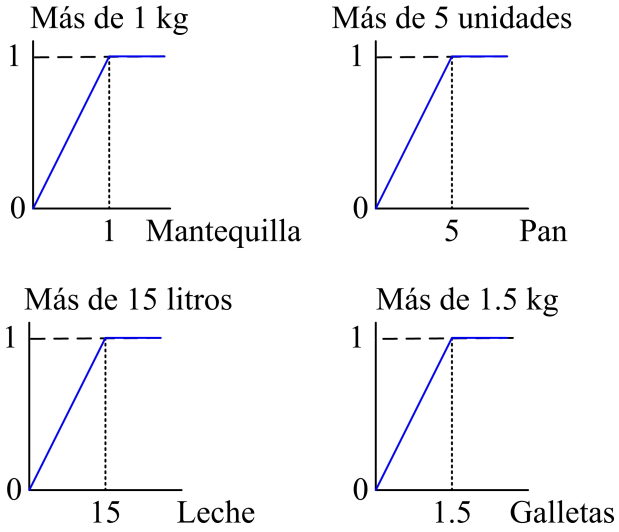

Figura 2.5: Conjuntos difusos para medir el grado de variación de las cantidades de los items de D_2.

Reglas	Soporte	Factor de Certeza
$(>, mantequilla) \rightarrow (>, leche)$	0.244	0.841
$(>, mantequilla) \rightarrow (>, galletas)$	0.227	0.74
$(>, leche) \rightarrow (>, galletas)$	0.244	0.857

Tabla 2.19: Dependencias graduales difusas en D_2 usando el Método 4.

> *"cuanto mayor (menor) hay una cantidad media (término lingüístico) de leche, mayor (menor) mucha (término lingüístico) cantidad de pan".*

- El método 3 extrae dependencias graduales difusas de la forma:

  > *"A más (menos) cantidad media (término lingüístico) de leche, más (menos) poca (término lingüístico) cantidad de galletas"*

  utilizando el modelo presentado en [Molina et al., 2007] para medir el grado de variación en conjunción con nuestro método para transformar una base de datos formada por bolsas en una transaccional difusa. Este tipo de reglas son útiles para obtener conocimiento sobre las tendencias existente en los datos

pero a distintos niveles o en distintas capas de análisis. Por ejemplo, una compañía podría estar interesada en las ventas de dos secciones distintas: ventas a pequeños clientes y ventas a grandes clientes. Para ello, antes del proceso de extracción, deberíamos usar una transformación del conjunto de bolsas (las ventas) usando dos términos lingüísticos: pequeñas y grandes, para distinguir los dos tipos de ventas que corresponderían a los dos tipos de clientes.

- El cuarto y último método es una propuesta que difiere del resto, pero que también usa el modelo descrito en [Molina et al., 2007]. Las reglas encontradas usando este método son de la forma $[>, leche] \rightarrow [<, mantequilla]$ y pueden interpretarse en nuestro ejemplo como

  > *"cuanto más (menos) difiera la cantidad de leche en más de quince litros, menor (mayor) será la diferencia de más de un kilogramo en la cantidad de mantequilla".*

  Este tipo de dependencias pueden ser útiles para obtener un conocimiento gradual sin tener que preocuparse en clasificar o dividir en conjuntos difusos el rango donde se mueven las cantidades asociadas a los items, pero sí teniendo en cuenta si la diferencia entre cantidades supera un umbral (ver Figura 2.5).

Observemos que con el método 4 podemos obtener (sin redundancia) dos tipos de reglas $[>, b_i] \rightarrow [>, b_j]$ y $[>, b_i] \rightarrow [<, b_j]$), mientras que con el método 3 el número de reglas es mayor, ya que es el doble del número de términos lingüísticos que hayamos usado para definir los nuevos items i_j. Por este hecho el método 3 tendrá una complejidad computacional mayor que el 4.

Los cuatro tipos de reglas, que podemos obtener según utilicemos cada uno de los métodos, se ajustan a distintas situaciones y proporcionan diferentes tipos de información. Cada tipo de regla se ha obtenido mediante una adecuada transformación del conjunto de bolsas a una base de datos transaccional difusa mediante la definición de conjuntos difusos apropiados. Puesto que estas transformaciones no son biyectivas, una vez efectuada la transformación no hay posibilidad de convertir un tipo de regla en otra, obteniendo, por tanto, cuatro clases distintas de reglas de asociación.

Algoritmos

Los métodos propuestos para extraer los distintos tipos de patrones en las bases de datos formadas por bolsas están basados en su transformación en transacciones difusas. Para ello el usuario debe dar un conjunto de etiquetas relacionadas con la cantidad de cada ítem representadas por conjuntos difusos. La segunda fase de todos ellos es la extracción de reglas de asociación difusas o dependencias graduales en las transacciones difusas. En nuestro caso particular, hemos utilizado las extensiones difusas del soporte y el factor de certeza dados en la sección 2.2. Para obtener reglas de asociación difusas podemos utilizar cualquier algoritmo ya existente en la literatura, destacando el presentado en [Delgado et al., 2003a].

Para el caso de las dependencias graduales, en [Molina et al., 2007] se trata el problema de la complejidad computacional de extraer los itemsets frecuentes en GT^D que es $O(n^2)$ en el número de objetos de la base de datos. Este problema puede resolverse considerando un número fijo de niveles equidistribuidos cuando se definen los conjuntos difusos. En este caso se podría conseguir un algoritmo para descubrir dependencias graduales cuya complejidad sería $n + k^{p+1}$ para itemsets de tamaño p donde n es el número de objetos en la base de datos.

Conclusiones

Las bolsas contienen más información que las transacciones normales. Para aprovechar dicha información hemos presentado varios métodos para extraer distintos tipos de patrones en este tipo de bases de datos, en particular hacemos uso de las reglas de asociación difusas y de las dependencias graduales tanto crisp como difusas.

Estos métodos se adaptan y pueden ser aplicados en un amplio rango de aplicaciones en las que se maneje información sobre frecuencias, cantidades junto con imprecisión, como hemos visto en el ejemplo 2.5 donde las bolsas difusas representaban de forma precisa la información sobre los términos y sus frecuencias de aparición en un conjunto de documentos.

La siguiente sección continua en la línea de la extracción de distintos tipos de conocimiento, esta vez basado principalmente en tres tipos de reglas: de excepción, anómalas y las que denominaremos como reglas dobles. También haremos uso del modelo formal para obtener un marco unificado para trabajar con los distintos tipos de reglas de asociación que será de utilidad para definir un algoritmo que ofrecerá la posibilidad de obtener diversos tipos de asociaciones con tan sólo cambiar el cuantificador-4ft que el usuario crea conveniente utilizar.

2.4. Búsqueda de otros Tipos de Conocimiento mediante Reglas de Asociación

Las reglas de asociación son una herramienta muy versátil que se acomoda a la extracción de múltiples tipos de información. En diversas situaciones se han modificado para dar lugar a nuevos tipos de reglas que ayudan a descubrir nuevo conocimiento en un conjunto de datos. Recientemente se han extendido para buscar nuevos modelos que encuentren información contenida en conjuntos de datos que no son frecuentes pero que contienen conocimiento de utilidad para el usuario. Ejemplos de estos nuevos tipos de reglas que se están desarrollando son las llamadas reglas infrecuentes, peculiares, de excepción o anómalas. La característica común de todas ellas es el bajo soporte de este tipo de reglas, de ahí que su desarrollo avance cada día hacia nuevos métodos y algoritmos para su extracción.

El principal objetivo de esta sección es hacer un breve recorrido sobre las distintas propuestas hechas hasta el momento en esta área, analizándolas desde un punto de vista práctico y semántico, y proveer nuevos procedimientos que ayuden al descubrimiento de este tipo de conocimiento.

El modelo formal desarrollado hasta el momento ofrece una buena herramienta para representar también este tipo de información, y mediante la definición de nuevos cuantificadores que extenderán a los ya conocidos cuantificadores-4ft podremos extraer nuevas reglas que capturen el conocimiento excepcional y anómalo en un conjunto de datos.

La sección se estructura como sigue: la sección 2.4.1 motiva el uso de este tipo de patrones y da una breve relación de los trabajos propuestos en el área. La sección 2.4.2 describe con más detalle los métodos propuestos para reglas de excepción y reglas anómalas, en las que nos centraremos en el resto de la sección. En las secciones 2.4.3 y 2.4.4 describiremos nuestras propuestas para extraer estos dos tipos de reglas, ofreciendo además su manejo y representación mediante el modelo formal que estamos desarrollando. Seguiremos con el estudio de las que denominaremos reglas dobles en la sección 2.4.5 y terminaremos con una discusión y algunas conclusiones.

2.4.1. Motivación

En muchas ocasiones la información que puede resultar de interés para el usuario es aquella que es inusual o inesperada en los datos. Esto hace replantearse la idea de extraer sólo aquellas reglas que sean frecuentes y confidentes. De esta

forma empiezan a surgir en los últimos años de la década de los 90 nuevos tipos de reglas que se centran en encontrar patrones infrecuentes en los datos.

Entre los trabajos más destacados en esta área podemos encontrar un gran número que estudian la búsqueda de excepciones o reglas de excepción pudiendo dividirse en dos grandes grupos según el tipo de búsqueda que empleen [Suzuki, 2004], [Taniar et al., 2008]:

- Directa. El método primero provee un conocimiento base normalmente en forma de reglas, y después se obtienen las reglas de excepción que se desvían o contradicen de alguna manera este conocimiento [Silberschatz and Tuzhilin, 1996], [Padmanabhan and Tuzhilin, 1998], [Liu et al., 1999a].
- Indirecta. En este tipo de métodos no hace falta dar ningún conocimiento previo para extraer las reglas de excepción [Suzuki and Shimura, 1996], [Taniar et al., 2008], [Hussain et al., 2000], [Suzuki and Zytkow, 2005], [Suzuki, 1997], [Liu et al., 1999b]. El objetivo se suele centrar en la búsqueda de pares de reglas que contienen la regla de excepción y su regla fuerte asociada.

Dentro de estos métodos se han seguido distintas aproximaciones para obtener un conjunto de reglas de excepción que sean de interés para el usuario. Entre las propuestas directas queremos destacar los enfoques subjetivos que permiten la participación del usuario para proveer el conocimiento base que se precisa para obtener en consecuencia la reglas que contradigan dicho conocimiento [Liu et al., 1999a], [Padmanabhan and Tuzhilin, 1998]. Esto no quiere decir que el usuario no pueda tomar parte en los métodos indirectos para ayudar a discernir las excepciones más interesantes o bien para definir los umbrales que sean necesarios en dicho proceso.

Si observamos con detenimiento el término *excepción* podemos ver que su significado nos dice que es algo que se aparta de la norma, es decir, *inusual*; o bien que contradice alguna regla, en otras palabras, *contradictorio*. De estos dos términos podemos ver con claridad que una *regla de excepción* deberá ser infrecuente (inusual) y contradictoria con respecto a otra regla que se denomina *regla fuerte* o *regla common sense*; de ahí que tengamos que tener bien una base de conocimiento o bien una regla fuerte que recoja un comportamiento 'normal' con el que contrarrestar la validez de la excepción.

Las reglas de excepción no son la única propuesta para extraer ciertos tipos de información inusual e interesante de las bases de datos. También podemos

encontrarnos con distintos trabajos que estudian cómo extraer reglas peculiares [Zhong et al., 2001], [Ohshima et al., 2007], reglas infrecuentes [Sim et al., 2008b] o reglas anómalas [Berzal et al., 2004], [Balderas et al., 2005].

Las *reglas peculiares* [Zhong et al., 2001] se extraen de una base de datos buscando cuáles son los datos relevantes entre aquellos que se consideran 'peculiares'. De forma poco precisa, se considera que un dato (una transacción) es peculiar si algunos de los atributos contienen algún valor peculiar. Un valor peculiar se reconocerá porque difiere bastante del resto de valores del atributo en la base de datos. Las reglas peculiares representan asociaciones entre estos datos peculiares. Este tipo de reglas se buscan entre aquellas que teniendo bajo soporte tengan valores altos para la confianza y alto cambio de soporte. En [Zhong et al., 2001] podemos encontrar un método orientado a los atributos para extraer reglas peculiares, y algunos trabajos recientes relacionados podemos verlos en [Ohshima et al., 2007].

Las *reglas infrecuentes* también han sido objeto de estudio puesto que a veces es interesante obtener información sobre patrones que no superan el umbral predefinido para el soporte. Este tipo de reglas se han utilizado principalmente en la detección de intrusiones junto con las excepciones [Yao et al., 2005b]. Se han seguido diversas técnicas para su extracción, entre ellas destacaremos su obtención mediante nuevas medidas que ayudan a discernir cuáles son las más interesantes: en [Sim et al., 2008b] modifican la medida conocida como *Lambda* para obtener las reglas infrecuentes más interesantes ayudándose de algunas técnicas de poda. En [Zhou and Yau, 2007b], [Zhou and Yau, 2007a] primero obtienen los items infrecuentes para después formar las reglas infrecuentes utilizando para ello las medidas de correlación, interés y el radio incremental de la probabilidad condicional entre dos items. En [Ding and Yau, 2009] se extraen las reglas infrecuentes usando una nueva estructura, en vez de nuevas medidas, llamada matriz de co-ocurrencia transaccional.

También hay otras propuestas como la de [Dong et al., 2007] que extraen los items infrecuentes para poder obtener otro tipos de reglas como por ejemplo reglas negativas (ver también [Sim and Indrawan, 2007], [Sim et al., 2008a]) que contienen este tipo de items.

En la siguiente sección repasaremos con más detenimiento las distintas propuestas dadas hasta el momento para la extracción y evaluación de las reglas de excepción y anómalas para poder compararlas después con nuestras propuestas desarrolladas en las secciones 2.4.3 y 2.4.4.

2.4.2. Otras Propuestas

Cuando el conocimiento que quiere obtenerse en una base de datos es de tipo inusual, inesperado, contradictorio y/o infrecuente, se han desarrollado modelos que extraen distintos tipos de conocimiento por medio de reglas de asociación adecuadas. Ejemplos de este tipo de reglas, hemos destacado anteriormente las reglas infrecuentes, peculiares, de excepción y anómalas. A continuación describiremos con más detalle aquellas propuestas que utilizan reglas de excepción y anómalas que serán los dos tipos que trataremos en el resto de esta sección.

Las propuestas relacionadas con el descubrimiento de reglas de excepción vimos que se podía dividir en técnicas directas e indirectas. Las técnicas directas generalmente son altamente subjetivas puesto que el conjunto de reglas con las que comparar suele depender del conjunto de creencias del usuario. De entre estos trabajos podemos destacar [Liu et al., 1999a], [Padmanabhan and Tuzhilin, 1998], [Silberschatz and Tuzhilin, 1996], en particular, el trabajo descrito [Liu et al., 1999a] lo describimos en la sección 2.1.3 donde las reglas de excepción son aquellas que son inesperadas por el usuario.

Las técnicas indirectas permiten obtener tanto el conjunto de reglas fuertes como las excepciones asociadas a ellas [Suzuki, 2004], [Yao et al., 2005a]. En [Duval et al., 2007] y [Taniar et al., 2008] se ofrece un resumen sobre las distintas propuestas que veremos con más detalle a continuación.

Entre los trabajos más importantes para la obtención de reglas de excepción están los desarrollados por Suzuki et al. y Hussain et al. que empezaron a desarrollarse a partir de 1996 continuando hasta nuestros días [Suzuki, 1996], [Suzuki, 2006].

La descripción del problema para extraer pares de reglas (common sense + excepción) en el caso de Suzuki et al. [Suzuki, 1997], [Suzuki, 2002] se plantea como encontrar pares de la forma siguiente

$$\begin{array}{rl} X \rightarrow y & \text{(regla common sense)} \\ X \wedge E \rightarrow y' & \text{(regla de excepción)} \end{array} \tag{2.23}$$

donde y e y' son dos valores distintos de un mismo ítem, y $X = x_1 \wedge \ldots \wedge x_p$, $E = e_1 \wedge \ldots \wedge e_q$ son dos itemsets (conjunciones de items). A la hora de obtener las reglas de excepción en lugar de usar las probabilidades asociadas a las anteriores reglas (conocidas como soporte y confianza), obtienen aquellos pares de reglas cuya probabilidad de tener los soportes y las confianzas oportunas mayor que los

umbrales establecidos sea mayor que $1-\delta$. Es decir, aquellos pares satisfaciendo[4] (2.24)$\sim$ (2.28).

$$P\{p(X) \geq \theta_1^s\} \geq 1-\delta \tag{2.24}$$

$$P\{p(y/X) \geq \theta_1^c\} \geq 1-\delta \tag{2.25}$$

$$P\{p(X,E) \geq \theta_2^s\} \geq 1-\delta \tag{2.26}$$

$$P\{p(y'|X,E) \geq \theta_2^c\} \geq 1-\delta \tag{2.27}$$

$$P\{p(y'|E) \leq \theta^I\} \geq 1-\delta \tag{2.28}$$

La ecuación (2.24) va asociada a la probabilidad real de que el soporte del ítem X, antecedente de la regla common sense (a partir de ahora notaremos por csr), supere el umbral θ_1^s que podríamos identificarlo como el $minsop$. La ecuación (2.25) representa la probabilidad real de que la confianza de la csr supere el umbral θ_1^s que de nuevo podemos identificarlo como la $minconf$. Las ecuaciones (2.26) y (2.27) son análogas a las anteriores para el soporte del antecedente de la regla de excepción (de ahora en adelante notada por exc) y para la confianza de la regla exc. La última ecuación (2.28) difiere del resto para obtener reglas de excepción que sean fiables y no provengan de una regla fuerte del tipo $E \rightarrow y'$, para ello se impone que la confianza de dicha regla no supere el umbral θ^I. A esta última regla se le llama regla de referencia y la notaremos por ref.

Para obtener dichas condiciones, ofrecen un método para estimar las probabilidades reales (2.24)$\sim$(2.28) usando aproximaciones normales de distribuciones multinomiales. En [Suzuki, 2004] podemos encontrar resumidos los distintos métodos seguidos para obtener los pares (csr, exc) del sistema descrito en (2.23).

En [Suzuki and Zytkow, 2005], [Suzuki, 2004] se considera una condición más asociada a la regla de referencia $E \rightarrow y'$ dada por la ecuación

$$p(E) \geq \theta_3^s \tag{2.29}$$

que impone que el itemsets E tenga un soporte mínimo.

A partir de esta formulación, llegan a un caso general donde se considera una terna de reglas, en vez de un par, donde una de ellas es una regla a la que llaman

[4] La notación está ligeramente cambiada de la ofrecida en [Suzuki, 1996] y [Suzuki, 2002] para comparar el método después con nuestra propuesta.

negativa. La terna que consideran se define como

$$t(x, y, \alpha, \beta, \gamma, \epsilon) = (x \rightarrow y, \alpha \not\rightarrow \beta, \gamma \rightarrow \epsilon) \tag{2.30}$$

donde las reglas del tipo $x \rightarrow y$ tienen asociadas las condiciones que llaman *generalidad* y *precisión* respectivamente:

$$\widehat{Pr}(x) \geq \theta_s \text{ (generalidad)} \quad \& \quad \widehat{Pr}(y|x) \geq \theta_c \text{ (precisión)} \tag{2.31}$$

y la regla negativa $\alpha \not\rightarrow \beta$ las condiciones:

$$\widehat{Pr}(\alpha) \geq \theta_s \quad \& \quad \widehat{Pr}(\beta|\alpha) \leq \theta_I \tag{2.32}$$

donde $\widehat{Pr}$ denota una estimación de la probabilidad real.

Con esta generalización del problema, según consideremos las variables $x, y, \alpha, \beta, \gamma, \epsilon$ obtendremos distintas estructuras que dan lugar a distintas definiciones de excepción. En particular, para el problema inicial que se planteaba [Suzuki, 1997], [Suzuki, 2002], la terna sería:

$$t(X, y, E, y', X \wedge E, y') = (X \rightarrow y, E \not\rightarrow y', X \wedge E \rightarrow y') \tag{2.33}$$

que denotan respectivamente la regla common sense, la regla de referencia y la regla excepcional (csr, ref, exc).

La propuesta indirecta de Liu et al. presentada en [Liu et al., 1999b] utiliza un método para identificar directamente las reglas de excepción usando desviaciones. El proceso que siguen tiene cuatro pasos bien diferenciados:

1. Se obtienen las reglas fuertes, filtrándolas para considerar aquellas que sean más fuertes o bien se escogen bajo cierto criterio como por ejemplo las que el usuario tenga más interés o aquellas que contengan los atributos más relevantes.

2. Para cada regla fuerte, se centran en sus atributos junto con sus clases para construir una tabla de contingencia que les ayude a calcular las desviaciones.

3. Según el valor de la desviación calculada (positiva, cero o negativa) y su magnitud se distinguen las reglas más interesantes y los atributos que ayudarán a conseguir las reglas de excepción.

4. Para conseguir las excepciones fiables, se toman los atributos de las desviaciones negativas encontradas y se buscan aquellas que tengan bajo soporte y alta confianza.

$X \rightarrow Y$	Regla common sense	(alto sop y alta $conf$)
$X \wedge E \rightarrow \neg Y$	Regla de excepción	(bajo sop y alta $conf$)
$E \rightarrow \neg Y$	Regla de referencia	(bajo sop y/o baja $conf$)

Tabla 2.20: Estructura esquematizada de las reglas de excepción dada en [Hussain et al., 2000].

Las reglas de excepción obtenidas siguiendo este proceso también llevan asociadas la regla common sense que sería la regla fuerte de la que proceden los atributos usados para calcular la excepción, y la regla de referencia, que puede obtenerse de la regla de excepción y que cumple las mismas propiedades que se imponen en la propuesta de Suzuki et al.

El método desarrollado en [Hussain et al., 2000] por Hussain et al. está basado en una nueva medida que estima el interés relativo de la regla de excepción a sus correspondientes csr y ref. Siguiendo la idea desarrollada por Suzuki et al., pero de forma más simplificada, las reglas de excepción se consideran como la terna de reglas de la Tabla 2.20. teniendo en cuenta que cuantos más atributos involucre el itemsets E, más reglas de referencia deberían imponerse, una por cada atributo en E.

En su propuesta, Hussain et al. utilizan como regla de referencia la regla $E \rightarrow Y$ imponiendo que sea del tipo csr, es decir, con altos valores para el soporte y la confianza, ya que imponer esta regla implica el cumplimiento de la regla de referencia $X \rightarrow Y$ [Hussain et al., 2000]. De esta forma, el algoritmo seguido se basa en obtener primero las dos csr y a partir de ellas obtener las reglas de excepción que tengan mayor valor en la medida que definen. Esta medida está basada en la diferencia de información relativa de exc con respecto a las dos csr. La Figura 2.6 esquematiza el método seguido en [Hussain et al., 2000].

La última propuesta que hemos encontrado para extraer reglas de excepción es la dada por Taniar et al. en [Taniar et al., 2008]. Esta propuesta difiere bastante de las anteriores ya que no sigue el esquema de encontrar pares o ternas de reglas que cumplen las propiedades expuestas anteriormente. En lugar de ello, basan la búsqueda de excepciones en encontrar aquellas reglas con bajo soporte y alta confianza del conjunto global de reglas obtenidas pudiendo ser éstas positivas o negativas. Además definen una medida para la excepcionabilidad que evalúa si la regla de excepción es fiable o no.

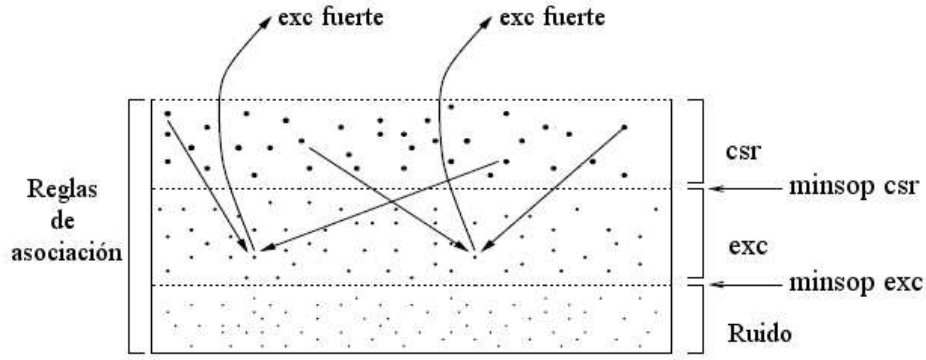

Figura 2.6: Esquema del método seguido en [Hussain et al., 2000].

Formalmente, para una regla fuerte positiva $X \rightarrow Y$, la posible excepción que puede encontrarse es $X \rightarrow \neg Y$ midiendo si supera un umbral impuesto al medir su excepcionabilidad y si además su soporte está comprendido en un intervalo $[s_1, s_2]$. Este intervalo sirve para impedir que la excepción no tenga soporte casi cero ni tampoco exceda el soporte mínimo. De forma análoga, para una regla fuerte negativa $X \rightarrow \neg Y$, la posible excepción sería $X \rightarrow Y$ cumpliendo las dos condiciones anteriores.

Observemos que esta última propuesta [Taniar et al., 2008] se distingue de las anteriores en que sólo encuentra el comportamiento inusual o contradictorio de una regla fuerte, mientras que las tres propuestas anteriores [Suzuki, 1997], [Suzuki, 2002], [Suzuki, 2004], [Liu et al., 1999b] y [Hussain et al., 2000] además de capturar dicho comportamiento, permiten saber qué o cuales atributos lo causan. Este es el papel del atributo o itemsets que hemos denominado E que al interaccionar con la regla common sense $X \rightarrow Y$ consigue contradecirla obteniendo otro valor del atributo Y, o de forma más general, se obtiene su contrario: $\neg Y$.

Así podríamos decir que las reglas de excepción extraídas con los procedimientos presentados en [Suzuki, 1997], [Suzuki, 2002], [Suzuki, 2004], [Liu et al., 1999b] y [Hussain et al., 2000] encuentran tanto el comportamiento excepcional (inusual y contradictorio) como el 'agente' que los causa.

Siguiendo la línea de trabajo detallada en el esquema de la Tabla 2.20, se pueden obtener distintos tipos de conocimiento imponiendo el cumplimiento de tres reglas (csr, exc, ref) que se ajusten al tipo de información que queramos extraer de una base de datos. Este es el caso de la propuesta dada por Berzal et al. en

[Berzal et al., 2004] y [Balderas et al., 2005] donde el nuevo tipo de conocimiento extraído se define como *reglas anómalas*.

De forma general, una regla anómala es una regla de asociación que aparece cuando eliminamos el efecto dominante de una regla del tipo common sense. En otras palabras, una regla anómala se cumple cuando la regla common sense falla [Berzal et al., 2004]. El conocimiento que se intenta capturar con este tipo de reglas es:

$$X \text{ implica fuertemente el cumplimiento de } Y,$$
$$\text{pero, en esos casos donde } X \text{ no implica } Y,$$
$$\text{entonces } X \text{ implica } A \text{ con una alta confianza,}$$

de forma breve: cuando ocurre X, entonces ocurre Y (usualmente) o A (inusualmente). El ítem A representará la anomalía.

El conocimiento dado por las reglas de excepción y las anómalas es en cierto modo semánticamente complementario. Si buscamos el 'agente' del comportamiento extraño o inusual, es conveniente buscar reglas de excepción, si por el contrario estamos más interesado en el comportamiento inusual, es mejor utilizar reglas anómalas.

Volviendo al tipo de representación anterior, para las reglas anómalas podríamos definir, siguiendo la definición de regla anómala en [Berzal et al., 2004], la terna $(csr, anom, ref)$ donde la regla common sense sería $X \rightarrow Y$, la regla anómala $X \wedge \neg Y \rightarrow A$ y para la regla de referencia imponen que $X \wedge Y \rightarrow \neg A$ sea fidedigna. Si viéramos esta última ref como una regla negativa en el sentido de Suzuki et al. podríamos decir que se cumple la regla de referencia $X \wedge Y \not\rightarrow A$.

El tipo de información extraído tanto por las reglas de excepción como por las anómalas es semánticamente distinto aunque ofrece similitudes en su formulación y pueden utilizarse métodos de extracción parecidos.

Las siguientes secciones analizan de forma detallada las similitudes y diferencias de estas propuestas desde un punto de vista semántico y formal, ofreciendo nuevos modelos para extraer reglas de excepción y anómalas utilizando para ello medidas de interés que no se han tenido aún en consideración en este campo y que proveen parte de la semántica recogida por la regla de referencia, pudiendo, por tanto, obviar su cumplimiento en algunos casos.

2.4.3. Nueva Propuesta para la Búsqueda de Reglas de Excepción

Las reglas de excepción capturan un tipo específico de conocimiento que puede interpretarse de forma sencilla como [Balderas et al., 2005]

$$X \text{ implica fuertemente el cumplimiento de } Y, \text{ (y no } E)$$
$$\text{pero, en conjunción con } E, X \text{ no implica } Y.$$

Por ejemplo [Suzuki, 2001], si X representa *antibióticos*, Y, *recuperación* y E, *estafilococos*, se podría encontrar la siguiente regla de excepción:

"con la ayuda de *antibióticos*, el paciente normalmente tiende a *recuperarse*,
a menos que aparezcan *estafilococos*",

en este caso la combinación de estafilococos con los antibióticos pueden causar la muerte. Este ejemplo muestra como la interacción del ítem E puede cambiar el comportamiento normal de la regla common sense, en este caso el valor de Y es la recuperación del paciente, mientras que $\neg Y$ (o y' en el enfoque de Suzuki et al.) es la muerte del paciente.

Mediante este ejemplo podemos ver claramente que la semántica de la regla de excepción está recogida con el par de reglas (csr, exc), mientras que la regla de referencia sirve para obtener reglas de excepción más fiables. Según la propuesta que manejemos, en la literatura podemos encontrar dos alternativas para definir la regla de referencia: la primera es $E \not\rightarrow \neg Y$ según el enfoque de Suzuki et al., pero en particular comprueban que $E \rightarrow \neg Y$ tenga bajo soporte y/o baja confianza, y la segunda es que $E \rightarrow Y$ sea una regla fuerte (alto soporte y confianza) para el enfoque de Hussain et al. [Hussain et al., 2000].

Nuestro modelo pertenece a las propuestas indirectas siguiendo el esquema de los enfoques de Suzuki et al. y de Hussain et al. Nuestra propuesta difiere de las anteriores en que nosotros sólo utilizaremos el par de reglas (csr, exc) para definir las excepciones y no impondremos la regla de referencia. La segunda diferencia radica en la medida utilizada para extraer las reglas de excepción, obteniendo un número menor de reglas de excepción sin necesidad de imponer la regla de referencia.

La primera razón que nos motiva a rechazar el uso de la regla de referencia es que no ofrece un enriquecimiento semántico al definir las reglas de excepción. La segunda proviene del hecho de que la regla de referencia propuesta con anterioridad

por Hussain et al. se sale del dominio del antecedente de la regla common sense, quedando por la propia definición de excepción, injustificado su uso. Además observemos que si obligamos que $E \rightarrow Y$ sea fuerte, estamos diciendo que $X \wedge E \rightarrow \neg Y$ es una excepción para dos *csr* al mismo tiempo: para $X \rightarrow Y$ y para $E \rightarrow Y$ por lo que el rol de 'agente' se intercambia, siendo E para la primera y X para esta última.

A continuación detallaremos nuestro modelo para la extracción de reglas de excepción usando el factor de certeza y analizaremos sus propiedades.

El Factor de Certeza para evaluar las Reglas de Excepción

El factor de certeza [Berzal et al., 2002] es una medida de interés que consigue extraer reglas de asociación muy fuertes (sección 2.1.8) y que tiene propiedades deseables para una medida de interés (sección 2.1.7). Cuando buscamos reglas de excepción tenemos la necesidad de diferenciar aquellas que son realmente fiables y que no surgen como consecuencia del cumplimiento de otras reglas fuertes. Este problema es el que se resolvía en las propuestas anteriores mediante la imposición de la regla de referencia. El principal inconveniente surge al definir una regla de referencia apropiada así como utilizar una medida de interés adecuada.

Nuestro modelo se basa en una reformulación del par (*csr*,*exc*) para obtener una semántica más precisa de regla de excepción. Además utilizaremos el soporte y el factor de certeza para obtener reglas de excepción que sean fiables, reduciendo por tanto, el número total de excepciones extraídas.

Definición 2.14. Sean X, Y y E itemsets arbitrarios en una base de datos D. Sea $D_X = \{t \in D : X \subset t\}$, es decir, D_X es el conjunto de transacciones en D que satisfacen X. Diremos que una *regla de excepción* es un par de reglas (*csr*, *exc*) de la forma:

- $X \rightarrow Y$ es muy fuerte en D
- $E \rightarrow \neg Y$ es fidedigna en D_X

donde la regla common sense (*csr*) es $X \rightarrow Y$ y la regla de excepción (*exc*), $E \rightarrow \neg Y$.

Mediante esta definición conseguimos dos objetivos muy importantes a la hora de extraer reglas de excepción:

1. Reducir la cantidad de pares (*csr*, *exc*) obtenidos.

2. Obtener reglas de excepción fiables.

En particular, para extraer las reglas muy fuertes utilizaremos el soporte y el factor de certeza (ver sección 2.1.8), y para medir que la regla de excepción *exc* sea fidedigna, utilizaremos el factor de certeza en vez de la confianza.

Observemos que cuando definimos la regla de excepción nos restringimos al dominio D_X puesto que queremos que la excepción sea cierta cuando nos limitamos al campo de acción del antecedente de la regla common sense. Si volvemos al ejemplo anterior, podemos ver claramente que la búsqueda de excepciones se centra en encontrar el agente E que al interactuar con X cambia el comportamiento usual de la regla de asociación *csr*.

Además nuestra definición para reglas de asociación puede verse como el par de reglas $(X \rightarrow Y,\ X \wedge E \rightarrow X \wedge \neg Y)$, pero esta última no está permitida en las definiciones usuales de regla de asociación puesto que el antecedente y el consecuente no son disjuntos. Sin embargo, mediante la restricción a D_X conseguimos que nuestra propuesta coincida con las anteriores cuando se utiliza como medida de interés la confianza. Si llamamos sop_X y Conf_X al soporte y la confianza obtenida en D_X, es fácil probar que $\text{Conf}(X \wedge E \rightarrow \neg Y) = \text{Conf}_X(E \rightarrow \neg Y)$:

$$\text{sop}_X(E) = \frac{|X \cap E|}{|X|} = \frac{|X \cap E| \,/\, |D|}{|X| \,/\, |D|} = \frac{\text{sop}(X \cup E)}{\text{sop}(X)} \tag{2.34}$$

$$\begin{aligned} \text{Conf}_X(E \rightarrow \neg Y) = \frac{\text{sop}_X(E \cup \neg Y)}{\text{sop}_X(E)} = \frac{|X \cap E \cap \neg Y| \,/\, |X|}{|X \cap E| \,/\, |X|} \\ = \frac{|X \cap E \cap \neg Y|}{|X \cap E|} = \text{Conf}(X \wedge E \rightarrow \neg Y) \end{aligned} \tag{2.35}$$

Recordemos algunas de las propiedades del factor de certeza que son de interés:

1. $\text{Conf}(\varphi \rightarrow \psi) = 1$ si y sólo si $FC(\varphi \rightarrow \psi) = 1$.

   Esta propiedad garantiza que el factor de certeza de una regla de asociación consigue su máximo valor, 1, si y sólo si la reglas es totalmente cierta [Delgado et al., 2003a].

2. Sea $\varphi \rightarrow \psi$ una regla de asociación con factor de certeza positivo. Entonces se cumple la siguiente igualdad [Berzal et al., 2002]

$$FC(\varphi \rightarrow \psi) = FC(\neg\psi \rightarrow \neg\varphi).$$

La confianza no satisface esta propiedad, y es muy interesantes ya que las reglas $\varphi \rightarrow \psi$ y $\neg\psi \rightarrow \neg\varphi$ representan el mismo tipo de conocimiento desde un punto de vista lógico.

En el capítulo 3.2 veremos como usando el factor de certeza conseguimos una reducción considerable del número de reglas extraídas y en consecuencia del número de reglas de excepción descubiertas.

A continuación adaptaremos el modelo formal que tenemos bajo estudio para la extracción de reglas de excepción.

Modelo formal para las Reglas de Excepción

Esta sección pretende ofrecer una visión formal tanto para la representación como la evaluación de las reglas de excepción tomando como herramienta el modelo formal. Mediante el uso de una adecuada tabla-4ft y adaptando los cuantificadores-4ft para extraer reglas de asociación en el conjunto D_X, presentaremos un modelo unificado que puede trabajar con pares de reglas, para las reglas de excepción, o con ternas como veremos en secciones posteriores para las reglas anómalas.

La ventaja de esta formulación radica en obtener un marco unificado para la extracción tanto de reglas de asociación, de excepción o anómalas, siguiendo el mismo proceso cambiando los cuantificadores-4ft necesarios en cada caso.

Consideraremos que X, Y y E son tres itemsets cualesquiera en una base de datos D, aunque en la práctica impondremos que E sea un sólo ítem para una mejor comprensión de la regla de excepción. Supongamos que la tabla-4ft asociada a la regla $X \approx Y$ es $\mathcal{M} = \langle a, b, c, d\rangle$. Definimos la tabla-4ft $4ft(E, Y, D_X) = \mathcal{M}_X^E = \langle e, f, g, h\rangle$ como

$\mathcal{M}_X^E$	Y	$\neg Y$	
E	e	f	
$\neg E$	g	h	
			$a+b$

donde e es el número de transacciones de D_X satisfaciendo Y y E; f el número de transacciones de D_X satisfaciendo E y no Y, etc. La suma de las cuatro frecuencias de $\mathcal{M}_X^E$ corresponde a la suma de a y b, i.e. $a = e + g$, $b = f + h$. Seguiremos usando n para el número total de transacciones de D.

Para adaptar los cuantificadores-$4ft$ de confianza y factor de certeza para validar la regla de excepción en D_X, es muy sencillo puesto que sólo tendremos

que tomar las frecuencias asociadas a los itemsets E y $\neg Y$ pero para la tabla-4ft $\mathcal{M}_X^E$. Usaremos el superíndice E en los cuantificadores-4ft para diferenciarlos de los definidos con anterioridad para $\mathcal{M}$.

Para la confianza recordemos que el cuantificador-4ft asociado es el implicacional, que para extraer reglas de excepción será

$$\Rightarrow_I^E (e, f, g, h) = \frac{f}{e + f} \tag{2.36}$$

Observemos que este cuantificador al involucrar un itemsets negado es distinto de $\Rightarrow_I$. En particular, $\Rightarrow_I^E (e, f, g, h) = \Rightarrow_I (f, e, h, g)$.

Estos cuantificadores-4ft, que llamaremos E-cuantificadores por brevedad, sólo tendrán asociada una condición del tipo siguiente:

$$\Rightarrow_I^E (e, f, g, h) \geq minconf \tag{2.37}$$

puesto que en principio no hay ninguna restricción para el soporte. Sin embargo si quisiéramos que el soporte de la regla exc estuviera contenido en un intervalo, como por ejemplo imponían algunas de las propuestas anteriores, sólo tendríamos que asociar también dicha condición que sería del tipo:

$$\frac{f}{e + f + g + h} = \frac{f}{a + b} \in [minEsop, maxEsop] \tag{2.38}$$

donde $minEsop$ y $maxEsop$ denotan los soportes mínimo y máximo que podría tener la regla exc. De nuevo, la condición estaría restringida a D_X, puesto que en el resto de la base de datos no importa si se cumple o no la excepción.

En nuestra propuesta, utilizaremos en lugar de la confianza, el cuantificador equivalente asociado al factor de certeza debido a sus buenas propiedades ya mencionadas en anteriores secciones. Además, gracias a la propiedad 1. del factor de certeza vista en la sección anterior, las reglas obtenidas son más fiables y no impondremos la condición para el soporte dada en (2.38).

El E-cuantificador asociado al factor de certeza se define como sigue:

$$\equiv_{FC}^E (e, f, g, h) = \begin{cases} \dfrac{\Rightarrow_I^E (e, f) - \mathrm{sop}_X(\neg Y)}{1 - \mathrm{sop}_X(\neg Y)} & \text{si } \Rightarrow_I^E (e, f) > \mathrm{sop}_X(\neg Y) \\ 0 & \text{si } \Rightarrow_I^E (e, f) = \mathrm{sop}_X(\neg Y) \\ \dfrac{\Rightarrow_I^E (e, f) - \mathrm{sop}_X(\neg Y)}{\mathrm{sop}_X(\neg Y)} & \text{si } \Rightarrow_I^E (e, f) < \mathrm{sop}_X(\neg Y) \end{cases} \tag{2.39}$$

que es equivalente a

$$\equiv_{FC}^{E}(e,f,g,h) = \begin{cases} \dfrac{fg-eh}{(e+f)(e+g)} & \text{si } fg > eh \\ 0 & \text{si } fg = eh \\ \dfrac{fg-eh}{(e+f)(f+h)} & \text{si } fg < eh. \end{cases} \tag{2.40}$$

El procedimiento a seguir para extraer reglas de excepción siguiendo nuestro modelo se podría resumir diciendo que en primer lugar utilizamos el cuantificador $\equiv_{FC}$ de forma *muy fuerte* como vimos en la definición 2.3 de la sección 2.1.8, y después imponemos que el E-cuantificador $\equiv_{FC}^{E}$ supere el umbral $minFC$ impuesto por el usuario.

2.4.4. Nuevas Propuestas para la Búsqueda de Reglas Anómalas

Las reglas anómalas fueron presentadas por primera vez por Berzal et al. en [Berzal et al., 2004] como una alternativa a las reglas de excepción intentando capturar un nuevo tipo de conocimiento en una base de datos.

De forma general, una regla anómala es una regla de asociación que sale a la superficie cuando se elimina el efecto dominante producido por una regla fuerte. En otras palabras, es una regla de asociación que se verifica cuando falla la regla common sense [Berzal et al., 2004]. Formalmente, una regla $X \rightsquigarrow A$ se dice que es una *regla anómala* si satisface las tres condiciones siguientes [Berzal et al., 2004], [Balderas et al., 2005]:

1. $X \rightarrow Y$ es una regla fuerte (frecuente y confidente)
2. $X \wedge \neg Y \rightarrow A$ es una regla confidente
3. $X \wedge Y \rightarrow \neg A$ es una regla confidente

Para enfatizar todos los consecuentes participantes, en [Balderas et al., 2005] se usa la notación $X \rightsquigarrow A|\neg Y$, que puede leerse como: "X está asociado con A cuando Y no está presente".

En esta definición está implícito el uso del soporte y la confianza para obtener las reglas anómalas. La semántica que intenta capturar este tipo de reglas es el siguiente:

X implica fuertemente Y,
pero en esos casos donde no se obtiene Y,
entonces X implica (de forma confidente) A.

En otras palabras: "cuando X, entonces se tiene Y (usualmente) o A (inusualmente)". Las reglas anómalas representan desviaciones homogéneas del comportamiento usual. Un ejemplo puede aclarar en qué situaciones es conveniente el uso de este tipo de reglas [Balderas et al., 2005]. Supongamos que en una base de datos médica obtenemos una regla fuerte del tipo

si síntomas-X entonces enfermedad-Y

y en aquellos casos donde la regla no se cumpla, podremos descubrir una anomalía interesante del tipo

si síntomas-X entonces enfermedad-A cuando no-enfermedad-Y

La diferencia esencial de las reglas anómalas con respecto a las de excepción es la existencia del itemsets E que hemos denominado el "agente" de la excepción en esta última, mientras que las reglas anómalas no buscan dicho agente, si no que pretenden encontrar el comportamiento causado mediante la imposición de que la mayoría de las excepciones correspondan al mismo consecuente A para poder considerarse una anomalía.

Si pensamos de nuevo en el esquema (csr, exc, ref), para las reglas anómalas podría verse como una terna (csr, $anom$, ref) donde la regla de referencia es $X \wedge Y \rightarrow \neg A$. En este caso, la regla de referencia sí viene definida para aquellas transacciones que satisfacen X ya que el conocimiento que queremos obtener está bajo la influencia de X y no importa lo que le ocurra a A fuera de su dominio.

De nuevo, la imposición de una regla de referencia ayuda a encontrar un conjunto de reglas anómalas que sea fiable, además de reducir su número. En esta línea se encuentra [Balderas et al., 2005] donde se continúa trabajando para encontrar reglas anómalas más fiables mediante algunos criterios que reducen el número de reglas anómalas extraídas. Entre estos criterios se encuentra el de proporcionar una medida de interés a cada regla para obtener un ranking de las reglas más fiables, en su caso se basa en la confianza de las reglas $anom$ y ref, es decir, cuanto mayor sea la confianza de $X \wedge \neg Y \rightarrow A$ y $X \wedge Y \rightarrow \neg A$ más fuerte es la anomalía.

Para resolver el problema de encontrar reglas anómalas más fiables y a la vez más fuertes a continuación presentamos nuestra propuesta usando el factor de certeza y siguiendo con la formulación presentada para las reglas de excepción.

Nueva Formulación para Reglas Anómalas

El modelo que proponemos para la extracción de reglas anómalas se basa en las mismas dos ideas que empleamos para las reglas de excepción:

1. Definir las reglas anómalas usando el dominio D_X.
2. Utilizar el factor de certeza en vez de la confianza para obtener reglas más fiables y reducir el número de reglas anómalas obtenido.

Antes de presentar nuestro modelo, analizaremos un poco más la formulación para reglas anómalas de [Berzal et al., 2004]. Los autores centran su búsqueda en las transacciones que, conteniendo X, no verifican la regla common sense $X \rightarrow Y$; y calculan la confianza de la regla anómala mediante la confianza de $X \wedge \neg Y \rightarrow A$ como sigue:

$$\mathrm{Conf}(X \rightsquigarrow A) = \frac{\mathrm{sop}(X \cup \neg Y \cup A)}{\mathrm{sop}(X \cup \neg Y)} \tag{2.41}$$

Además imponen que la confianza de la regla de referencia

$$\mathrm{Conf}(X \wedge Y \rightarrow \neg A) = \frac{\mathrm{sop}(X \cup Y \cup \neg A)}{\mathrm{sop}(X \cup Y)} = \frac{\mathrm{sop}(X \cup Y) - \mathrm{sop}(X \cup Y \cup A)}{\mathrm{sop}(X \cup Y)} \tag{2.42}$$

sea mayor que el umbral $minconf$.

Veamos con más detenimiento qué nos dice esta condición. Si observamos la Figura 2.7 podemos ver que fijando el soporte de $X \cup Y \cup A$, la confianza de la regla de referencia es monótona creciente en $\mathrm{sop}(X \cup Y)$, es decir, que crece cuanto mayor es el soporte de la regla common sense. Además este crecimiento es muy fuerte cuando se supera el valor impuesto para el $minsop$ (en el caso de la gráfica, $minsop = 0.01$). Si por el contrario fijamos $\mathrm{sop}(X \cup Y)$, la confianza depende linealmente de $\mathrm{sop}(X \cup Y \cup A)$ y decrece hasta 0 cuando $\mathrm{sop}(X \cup Y \cup A) = minsop$.

Por tanto, el crecimiento de $\mathrm{Conf}(X \wedge Y \rightarrow \neg A)$ es mayor cuanto mayor sea $\mathrm{Sop}(X \rightarrow Y)$. Además está más próximo a 1 cuando el anterior soporte ha sobrepasado el umbral $minsop$. Esto nos lleva a que la condición impuesta por la regla de referencia es bastante débil puesto que $\mathrm{Sop}(X \rightarrow Y) \geq minsop$ siempre se cumple por las condiciones asociadas a la regla common sense. También es cierto que cuanto menor sea $\mathrm{sop}(X \cup Y \cup A)$, la confianza de la regla de referencia será menor.

Por este motivo proponemos una formulación distinta para obtener reglas anómalas que difiere en la anterior en la regla de referencia, que después probaremos que es una condición más fuerte que la dada en [Berzal et al., 2004] y [Balderas et al., 2005].

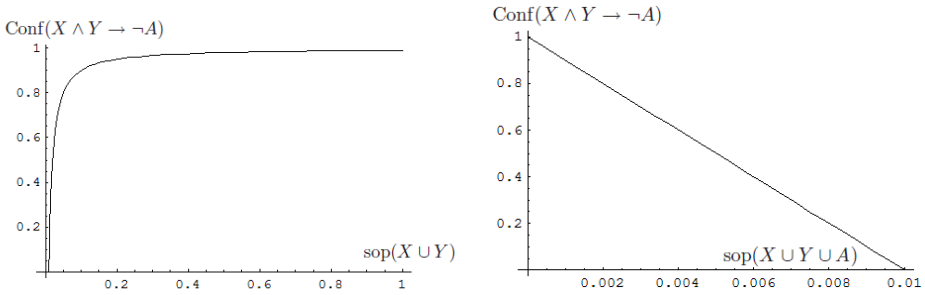

Figura 2.7: Comportamiento de la confianza de la regla de referencia frente a los soportes de $X \cup Y$ y $X \cup Y \cup A$.

Definición 2.15. Sean X, Y dos itemsets arbitrarios y A un ítem arbitrario. Definiremos una regla anómala por la terna (csr, $anom$, ref) cumpliendo las siguientes condiciones:

1. $X \rightarrow Y$ es una regla fuerte (csr, frecuente y confidente).
2. $\neg Y \rightarrow A$ es una regla confidente en D_X ($anom$).
3. $A \rightarrow \neg Y$ es una regla confidente en D_X (ref).

Esta definición lleva el uso implícito del soporte y la confianza como medidas para extraer las reglas anómalas (como en la definición propuesta en [Berzal et al., 2004]), usando para ello los umbrales de $minsop$ y $minconf$.

Si comparamos nuestra formulación con la de Berzal et al., la nuestra es equivalente desde un punto de vista lógico formal si $anom$ y ref las definimos en D_X puesto que $A \rightarrow \neg Y$ es equivalente a $\neg\neg Y \rightarrow \neg A \equiv Y \rightarrow \neg A$. Además si notamos por sop_X, Sop_X y Conf_X respectivamente al soporte de un itemsets, de una regla y la confianza de una regla en D_X tenemos que:

$$\begin{aligned}
\text{sop}_X(A) &= \frac{|A_X|}{|D_X|} = \frac{|X \cap A|}{|X|} = \frac{|X \cap A| \,/\, |D|}{|X| \,/\, |D|} = \frac{\text{sop}(X \cup A)}{\text{sop}(X)} \\
\text{sop}_X(A \rightarrow B) &= \frac{\text{sop}_X(A \cup B)}{|D_X|} = \frac{\text{sop}(X \cup A \cup B)}{\text{sop}(X)} = \frac{|X \cap A \cap B|}{|X|} \\
\text{Conf}_X(A \rightarrow B) &= \frac{\text{sop}_X(A \cup B)}{\text{sop}_X(A)} = \frac{|X \cap A \cap B| \,/\, |X|}{|X \cap A| \,/\, |X|} = \frac{|X \cap A \cap B|}{|X \cap A|} \\
&= \text{Conf}(X \wedge A \rightarrow B)
\end{aligned}$$

donde $A_X = \{t \in A : X \subset t\}$. Y usando las anteriores igualdades es fácil probar que:

$$\begin{aligned}
\text{Conf}(X \wedge \neg Y \rightarrow A) &= \text{Conf}_X(\neg Y \rightarrow A) \\
\text{Conf}(X \wedge Y \rightarrow \neg A) &= \text{Conf}_X(Y \rightarrow \neg A) \\
\text{Conf}(X \wedge A \rightarrow \neg Y) &= \text{Conf}_X(A \rightarrow \neg Y).
\end{aligned}$$

Queremos destacar que la confianza asociada a nuestra regla de referencia tiene un comportamiento parecido al ofrecido para la de Berzal et al. en la Figura 2.7 pero en la primera gráfica el parámetro que influye es $\text{sop}(X \cup A)$ sobre el que "a priori" no tenemos ninguna condición asociada.

Veamos mediante un ejemplo como nuestra formulación es más fuerte que la ofrecida por Berzal et al. [Berzal et al., 2004], introduciendo un punto de vista semántico ligeramente distinto al cambiar la regla de referencia.

X	Y	A_1	$\cdots$
X	Y	A_1	$\cdots$
X	Y	A_2	$\cdots$
X	Y	A_2	$\cdots$
X	Y	A	$\cdots$
X	Y	A	$\cdots$
X	Y	A	$\cdots$
X	Y	A	$\cdots$
X	Y'	A	$\cdots$
X	Y'	A	$\cdots$
	$\cdots$		

X	Y	A_1	$\cdots$
X	Y	A_1	$\cdots$
X	Y	A_2	$\cdots$
X	Y	A_2	$\cdots$
X	Y	A_3	$\cdots$
X	Y	A	$\cdots$
X	Y	A	$\cdots$
X	Y	A	$\cdots$
X	Y'	A	$\cdots$
X	Y'	A	$\cdots$
	$\cdots$		

X	Y	A_1	$\cdots$
X	Y	A_1	$\cdots$
X	Y	A_2	$\cdots$
X	Y	A_2	$\cdots$
X	Y	A_3	$\cdots$
X	Y	A_3	$\cdots$
X	Y	A	$\cdots$
X	Y	A	$\cdots$
X	Y'	A	$\cdots$
X	Y'	A	$\cdots$
	$\cdots$		

Tabla 2.21: Conjuntos de datos D1, D2, D3.

En los conjuntos de datos D1, D2 y D3 representados en la Tabla 2.21 sólo hemos considerado aquellas transacciones que satisfacen X. En todos ellos la fuerza de la regla common sense y la confianza de la anomalía son la misma. En particular, $\text{Conf}(X \wedge \neg Y \rightarrow A) = 1$.

Si medimos la confianza de $X \wedge Y \rightarrow \neg A$ (ref en la propuesta de Berzal et al.) es bastante alta en todas ellas, superando el valor de 0.5, mientras que la confianza

de $X \wedge A \rightarrow \neg Y$ (ref en nuestra propuesta) no supera dicho valor hasta el conjunto D3. La Tabla 2.22 muestra la confianza usando ambas propuestas.

D	Conf$(X \wedge Y \rightarrow \neg A)$	Conf$(X \wedge A \rightarrow \neg Y)$
1	0.5	0.333
2	0.625	0.4
3	0.75	0.5

Tabla 2.22: Confianza de las reglas $X \wedge Y \rightarrow \neg A$ y $X \wedge A \rightarrow \neg Y$ en los conjuntos de datos D1, D2 y D3.

Resumiendo, nuestra condición es más estricta bajo la condición que marca la proposición 2.6 y la semántica asociada es que el porcentaje de transacciones conteniendo X y A y no Y entre aquellas que contienen X y A tiene que ser mayor que el umbral para la confianza. Aunque nuestra propuesta es equivalente desde la lógica formal a la propuesta de Berzal et al., en ella no tenemos el problema de que la confianza de la regla $X \wedge Y \rightarrow \neg A$ se vea afectada cuando se incrementa el soporte de $X \rightarrow Y$.

La siguiente proposición ofrece una comparación del comportamiento de la regla de referencia de ambas propuestas.

Proposición 2.6. *Sean X, Y y A itemsets arbitrarios. Se cumple la siguiente desigualdad*

$$\text{Conf}(X \wedge A \rightarrow \neg Y) \leq \text{Conf}(X \wedge Y \rightarrow \neg A) \tag{2.43}$$

si, y sólo si,

$$\text{sop}(X \cup A) \leq \text{sop}(X \cup Y).$$

Demostración. Por definición de la confianza deberemos probar:

$$\frac{\text{sop}(X \cup A \cup \neg Y)}{\text{sop}(X \cup A)} \leq \frac{\text{sop}(X \cup Y \cup \neg A)}{\text{sop}(X \cup Y)}$$

y sabemos que

$$\text{sop}(X \cup A) = \text{sop}(X \cup A \cup \neg Y) + \text{sop}(X \cup A \cup Y)$$
$$\text{sop}(X \cup Y) = \text{sop}(X \cup Y \cup \neg A) + \text{sop}(X \cup A \cup Y)$$

Para simplificar los cálculos, llamaremos $j = \text{sop}(X \cup A \cup \neg Y)$, $k = \text{sop}(X \cup Y \cup \neg A)$ e $i = \text{sop}(X \cup A \cup Y)$. Con esta notación tenemos que $\text{sop}(X \cup A) = i + j$ y

$\text{sop}(X \cup Y) = i + k$. Luego, deberemos probar cuándo se cumple la siguiente desigualdad:

$$\frac{j}{i+j} \leq \frac{k}{i+k}.$$

Como j, k e i son ≥ 0, la anterior desigualdad es equivalente a:

$$\begin{aligned} j(i+k) &\leq k(i+j) \\ ji + jk &\leq ki + kj \\ ji &\leq ki \\ j &\leq k \end{aligned}$$

donde hemos supuesto que $i = \text{sop}(X \cup A \cup Y)$ es positivo ($i \neq 0$). Si $i = 0$ tendríamos una situación trivial ya que $j/j = 1 \leq 1 = k/k$ siempre se cumple.

Por tanto, la desigualdad de la ecuación (2.43) se verifica si, y sólo si $\text{sop}(X \cup A \cup \neg Y) \leq \text{sop}(X \cup Y \cup \neg A)$ que es equivalente a imponer que $\text{sop}(X \cup A) \leq \text{sop}(X \cup Y)$, ya que el cumplimiento de $j \leq k$ es equivalente al cumplimiento de $i+j \leq i+k$. ■

Esta proposición proporciona una relación entre las confianzas de las reglas de referencia de las dos propuestas, probando que la nuestra es más restrictiva que la ofrecida en [Berzal et al., 2004].

En general, la desigualdad (2.43) es cierta cuando el soporte de $X \cup Y$ es mayor que el de $X \cup A$. Si suponemos lo contrario, es decir, si $\text{sop}(X \cup A)$ fuera mayor que $\text{sop}(X \cup Y)$ esto implicaría que la confianza de la regla $X \rightarrow A$ es mayor que la de la regla $X \rightarrow Y$,

$$\frac{\text{sop}(X \cup Y)}{\text{sop}(X)} = \frac{|X \cap Y|}{|X|} < \frac{|X \cap A|}{|X|} = \frac{\text{sop}(X \cup A)}{\text{sop}(X)}$$

esto es,

$$\text{Conf}(X \rightarrow Y) < \text{Conf}(X \rightarrow A)$$

luego $X \rightarrow A$ sería más fuerte que $X \rightarrow Y$. Por tanto, si el soporte de $X \cup A$ es mayor que el de $X \cup Y$ estaríamos buscando las reglas anómalas asociadas a la regla common sense $X \rightarrow A$ en lugar de $X \rightarrow Y$. Por tanto, la condición dada en la proposición 2.6 se cumplirá siempre.

El Factor de Certeza para Evaluar las Reglas Anómalas

El uso de la confianza para extraer reglas fiables no es muy aconsejable debido a los problemas que ya comentamos en el primer capítulo. En particular, cuando queramos obtener reglas anómalas fiables, será mejor utilizar otro tipo de medida

de interés. En nuestro caso, el factor de certeza será la medida elegida, quedando la obtención de las reglas anómalas definida por la terna $(csr, anom, ref)$ cumpliendo las tres siguientes propiedades:

1. $X \rightarrow Y$ es una regla muy fuerte (csr, frecuente y fidedigna).
2. $\neg Y \rightarrow A$ es una regla fidedigna en D_X ($anom$).
3. $A \rightarrow \neg Y$ es una regla fidedigna en D_X (ref).

En esta nueva formulación viene implícito el uso del soporte y el factor de certeza así como los umbrales asociados $minsop$ y $minFC$.

Observemos que usando el factor de certeza, las equivalencias anteriores cuando usábamos la confianza en D y en D_X no son ciertas. Esto es debido a que para calcular el factor de certeza en D, utilizamos el soporte del consecuente en D, mientras que en D_X el soporte se calcula en D_X. Con otras palabras,

$$\begin{aligned} FC(X \wedge \neg Y \rightarrow X \wedge A) &\neq FC_X(\neg Y \rightarrow A) \\ FC(X \wedge A \rightarrow X \wedge \neg Y) &\neq FC_X(A \rightarrow \neg Y) \end{aligned} \tag{2.44}$$

puesto que

$$\text{sop}(X \wedge A) = \frac{|X \cap A|}{|D|} \neq \frac{|X \cap A|}{|X|} = \text{sop}_X(A). \tag{2.45}$$

A continuación del mismo modo que hicimos para las reglas de excepción, adaptaremos los cuantificadores-$4ft$ para las reglas anómalas.

Modelo Formal para las Reglas Anómalas

La formulación empleada para definir las reglas anómalas se asemeja al de las reglas de excepción, por lo que los cuantificadores-$4ft$ asociadas a ambos tipos de reglas guardarán ciertas relaciones entre sí.

En [Delgado et al., 2008b] podemos encontrar un estudio preliminar del modelo formal para la representación y la evaluación de las reglas de excepción y las anómalas. Pero aquí presentamos un estudio más detallado y ligeramente distinto al presentado allí, resaltando la relación entre los cuantificadores-$4ft$ de los dos tipos de reglas estudiados.

Sean X, Y, y A los itemsets involucrados en la formulación de las reglas anómalas. Del mismo modo que ocurría para las reglas de excepción, la tabla-$4ft$ asociada es $\mathcal{M}_X^A = 4ft(A, Y, D_X) = \langle i, j, k, l \rangle$ donde el papel de E ahora lo desempeña el itemsets A (ver Tabla 2.23). En dicha tabla i representa el número

$\mathcal{M}_X^A$	Y	$\neg Y$	
A	i	j	
$\neg A$	k	l	
			$a+b$

Tabla 2.23: Tabla-4ft para los itemsets A e Y en D_X.

de transacciones en D satisfaciendo X, Y y A; j el número de transacciones satisfaciendo X, $\neg Y$ y A y así sucesivamente con k y l. La suma de estas frecuencias corresponden a las frecuencias a y b vistas en la sección 1.2, i.e. $a = i+k$, $b = j+l$. Seguiremos notando por n al numero total de transacciones de D.

De nuevo adaptaremos los cuantificadores-4ft para la evaluación de reglas anómalas. En particular, sólo deberemos definir dos nuevos cuantificadores que son los asociados a la regla denominada *anom*, puesto que para la regla *csr* y para la ref podemos reutilizar los definidos para las reglas de excepción.

Llamaremos A-cuantificador, para distinguirlo de los anteriores, cuando el cuantificador use las frecuencias de la tabla-4ft $\mathcal{M}_X^A$. Para las dos propuestas anteriores, definimos los dos A-cuantificadores siguientes:

$$\Rightarrow_I^A (i,j,k,l) = \frac{j}{j+l} \tag{2.46}$$

para la primera propuesta que exige que $\neg Y \rightarrow A$ sea confidente en D_X, y

$$\equiv_{FC}^A (i,j,k,l) = \begin{cases} \dfrac{jk-il}{(j+l)(k+l)} & \text{if } jk > il \\ 0 & \text{if } jk = il \\ \dfrac{jk-il}{(i+j)(j+l)} & \text{if } jk < il. \end{cases} \tag{2.47}$$

para la segunda propuesta donde se exige que la regla anómala sea fidedigna.

La condición asociada a la regla common sense será idéntica que para el caso de las excepciones usando el cuantificador $\Rightarrow_I (a,b,c,d)$ de forma fuerte o el cuantificador $\equiv_{FC} (a,b,c,d)$ de forma muy fuerte, según la propuesta.

Para la regla de referencia, observemos que su formulación coincide con la de la regla de excepción *exc* donde se intercambia E por A. Por tanto, podemos utilizar los E-cuantificadores definidos anteriormente pero usando las frecuencias de $\mathcal{M}_X^A$, es decir, los cuantificadores $\Rightarrow_I^E (i,j,k,l)$ y $\equiv_{FC}^E (i,j,k,l)$ según sea la regla de referencia confidente o fidedigna respectivamente.

Siguiendo la notación del modelo formal, podemos ver que la proposición 2.6 puede reformularse como se muestra a continuación, siendo su demostración la misma ya que utilizamos justamente una notación que concuerda con la del modelo.

Proposición 2.7. *Sea* $\mathcal{M}_X^A = 4ft(A, Y, D_X) = \langle i, j, k, l\rangle$ *la tabla-*$4ft$ *asociada a los itemsets* X, Y *y* A*. Entonces, la desigualdad siguiente*

$$\Rightarrow_I^E (i, j, k, l) \leq \Rightarrow_I^E (i, k, j, l) \tag{2.48}$$

se cumple si, y sólo si

$$j \leq k.$$

Su equivalencia con la proposición 2.6, es inmediata si observamos que $\text{Conf}_X(A \rightarrow \neg Y) = \Rightarrow_I^E (i, j, k, l)$, y $\Rightarrow_I^E (i, k, j, l) = \dfrac{k}{i+k} = \text{Conf}_X(Y \rightarrow \neg A)$.

Resumiendo, para obtener reglas anómalas usando la definición 2.15 en la que se tiene sólo en consideración la confianza, los cuantificadores involucrados son:

- $\Rightarrow_I (a, b, c, d)$ usado de forma fuerte, i.e. $\Rightarrow_I (a, b, c, d) \geq minconf$ y $\dfrac{a}{n} > minsop$,
- $\Rightarrow_I^A (i, j, k, l) \geq minconf$, y
- $\Rightarrow_I^E (i, j, k, l) \geq minconf$.

Por el contrario si queremos utilizar la segunda propuesta para obtener reglas anómalas más fiables en la que se ve involucrado el factor de certeza:

- $\equiv_{FC} (a, b, c, d)$ usado de forma muy fuerte, i.e. $\equiv_{FC} (a, b, c, d) \geq minFC$, $\dfrac{a}{n} > minsop$ y $\dfrac{d}{n}$,
- $\equiv_{FC}^A (i, j, k, l) \geq minFC$, y
- $\equiv_{FC}^E (i, j, k, l) \geq minFC$.

A continuación presentaremos un nuevo tipo de regla de asociación a la que llamaremos *regla doble*. Este tipo de regla junto con la extracción de sus reglas de excepción y anómalas asociadas, nos describirá de una forma más precisa y fiable la relación entre dos itemsets.

2.4.5. Describiendo la Relación Existente entre los Itemsets: Reglas Dobles

Las reglas de asociación tratan de describir el tipo de relación existente cuando se encuentran ocurrencias conjuntas de items de diversa índole. Según la medida que usemos, podremos percibir si la relación cumple unas propiedades u otras. Pero en la mayoría de los casos la medida no es capaz de describir de forma fiable dicha relación porque los valores obtenidos no alcanzan su máximo. Por esta razón quedan muchos flecos cuando se intenta describir la relación entre dos o más itemsets.

El propósito de esta sección es ofrecer un procedimiento para describir de la forma más completa posible la relación entre un conjunto de items mediante algunas de las herramientas descritas anteriormente. La idea clave en el proceso es la extracción de las que llamaremos *reglas dobles* junto con las reglas anómalas y de excepción asociadas.

A veces, cuando se efectúa una búsqueda de reglas de asociación, se encuentran que dos itemsets están muy relacionados y existe poca diferencia si escogemos un sentido u otro de la relación. Cuando nos encontremos en dicha situación, será útil obtener no sólo en qué sentido es más fuerte la relación entre ambos, si no los dos sentidos de la regla de asociación junto con las medidas asociadas para cada sentido, y de este modo estudiar con más profundidad la naturaleza de estos resultados. Para ello definiremos el concepto de *regla doble* para representar este tipo de reglas "bidireccionales".

Por tanto, una regla doble será una regla que represente una relación fuerte en las dos direcciones posibles: en $X \rightarrow Y$ y en $Y \rightarrow X$. Veamos un ejemplo de este tipo de situaciones.

Ejemplo 2.6. Imaginemos que tenemos una base de datos que almacena información sobre los animales vertebrados y sus características en un parque nacional. Una de las reglas fuertes que podríamos extraer es:

si el animal vuela, entonces es un pájaro.

Pero esta regla también puede extraerse de forma fuerte en la otra dirección, i.e. la regla

si el animal es un pájaro, entonces vuela

es también fuerte.

En estas situaciones, será conveniente no desechar una de las direcciones escogiendo la de mayor valor, si no, obtener la regla en ambas direcciones siempre que se superen unos umbrales mínimos establecidos por el usuario. Siguiendo este criterio definiremos las reglas dobles.

Definición 2.16. Una regla de asociación $X \rightarrow Y$ se dice que es *doble y fuerte* si las reglas $X \rightarrow Y$ e $Y \rightarrow X$ son fuertes.

De ahora en adelante utilizaremos $\leftrightarrow$ para denotar a las reglas dobles. En particular llamaremos antecedente y consecuente de una regla doble $X \leftrightarrow Y$ a X e Y respectivamente, donde la confianza (u otra medida de interés como el factor de certeza) de la regla de $X \rightarrow Y$ será mayor o igual que la de $Y \rightarrow X$.

De acuerdo a esta definición, proponemos su análoga para un cuantificador-$4ft$.

Definición 2.17. Sean X, Y dos itemsets en una base de datos D formada por n transacciones. Sea $\mathcal{M} = 4ft(X, Y, D) = \langle a, b, c, d \rangle$ la tabla-$4ft$ asociada. Diremos que un cuantificador-$4ft \approx$ se utiliza de forma *doblemente fuerte* si se le imponen las siguientes condiciones:

$$\approx (a, b, c, d),\ \approx (a, c, b, d) \geq p \ \wedge\ \frac{a}{n} \geq minsop$$

donde $0 < p < 1$ y $0 < minsop < 1$.

Queremos hacer notar que el soporte de las reglas $X \rightarrow Y$ e $Y \rightarrow X$ coinciden, y que p en la definición anterior es el umbral definido para el cuantificador. Por ejemplo en el caso de utilizar el cuantificador-$4ft \Rightarrow_I$, p es justamente $minconf$ y en el caso de $\equiv_{FC}$, p es $minFC$.

Recordemos que un cuantificador-$4ft$ es simétrico [Hájek and Havránek, 1978] si $\approx\ (a, b, c, d) = \approx\ (a, c, b, d)$. Para este tipo especial de cuantificador-$4ft$ se cumple el siguiente corolario.

Corolario 2.3. *Un cuantificador-$4ft$ simétrico $\approx$ se utiliza de forma* doblemente fuerte *si satisface las siguientes condiciones:*

$$\approx (a, b, c, d) \geq p \ \wedge\ \frac{a}{n} \geq minsop$$

donde $0 < p < 1$ *y* $0 < minsop < 1$.

Esta propiedad reduce el número de imposiciones para extraer reglas dobles si utilizamos cuantificadores-$4ft$ simétricos.

La idea de regla doble y fuerte captura una relación muy estrecha entre dos conjuntos de items. Podría verse como una especie de equivalencia entre los dos itemsets en el sentido de que son en cierta medida intercambiables.

En el ejemplo anterior podemos ver que la regla doble *pájaro* $\leftrightarrow$ *volar* no es 100 % cierta ya que hay casos en los que no se cumple, por ejemplo un pingüino es un pájaro que no vuela. En las siguientes secciones proponemos cómo obtener las reglas de excepción y las anómalas asociadas a una regla doble, explicando las posibles situaciones y su semántica asociada.

Reglas de Excepción asociadas a Reglas Dobles

Las reglas dobles pueden verse como dos reglas de asociación simples, por ello pueden tener asociadas distintas reglas de excepción en cada uno de los sentidos de la regla. Estas reglas de excepción mostrarán con más detalle cuáles son los "agentes" que interfieren en la relación entre el antecedente y el consecuente de la regla doble. Veamos el siguiente ejemplo, continuación del anterior.

Ejemplo 2.7. En el ambiente del ejemplo 2.6 podemos descubrir dos tipos de reglas de excepción fijada la regla common sense *pájaros* $\leftrightarrow$ *volar*. El primer tipo de regla de excepción contiene las excepciones asociadas al antecedente de la regla doble:

si el animal es un pájaro, entonces vuela,

excepto los pingüinos y las avestruces

y el segundo tipo está asociado al consecuente, como por ejemplo:

si el animal vuela, entonces es un pájaro, *excepto* los murciélagos.

Como se muestra en el ejemplo, las reglas de excepción pueden estar asociadas a los dos sentidos de la regla doble. A continuación analizamos los distintos casos poniendo un énfasis especial en algunos que son más interesantes. Las distintas alternativas son:

1. No encontrar reglas de excepción asociadas a la regla doble.
2. Encontrar reglas de excepción en un sólo sentido de la regla doble, obteniendo uno de los dos esquemas siguientes:

bien	$X \leftrightarrow Y$	Regla doble y fuerte
	$X \wedge E \rightarrow \neg Y$	Regla confidente/fidedigna

o bien	$X \leftrightarrow Y$	Regla doble y fuerte
	$Y \wedge E \rightarrow \neg X$	Regla confidente/fidedigna

3. Encontrar reglas de excepción en ambos sentidos de la regla doble:

$X \leftrightarrow Y$	Regla doble y fuerte
$X \wedge E \rightarrow \neg Y$	Regla confidente/fidedigna
$Y \wedge E' \rightarrow \neg X$	Regla confidente/fidedigna

En este caso, podemos encontrarnos con dos situaciones:

a) La regla doble tiene distintas reglas de excepción en cada dirección, es decir, los items involucrados en ellas son distintos, i.e. $E \neq E'$.

b) Las reglas de excepción involucran el mismo ítem en los dos sentidos de la regla doble, i.e. tenemos que $E = E'$:

$X \leftrightarrow Y$	Regla doble y fuerte
$X \wedge E \rightarrow \neg Y$	Regla confidente/fidedigna
$Y \wedge E \rightarrow \neg X$	Regla confidente/fidedigna

Para ilustrar que este último caso puede ocurrir, consideramos la base de datos mostrada en la Tabla 2.24 que contiene doce transacciones. En ella tenemos que $\text{sop}(X \leftrightarrow Y) \simeq 0.583$, $\text{Conf}(X \rightarrow Y) \simeq 0.78$, y $\text{Conf}(Y \rightarrow X) \simeq 0.78$, lo que demuestra que $X \leftrightarrow Y$ es una regla doble y fuerte si imponemos el umbral $minsop = 0.5$ para el soporte y, para la confianza, $minconf = 0.6$. Además se cumple que $\text{Conf}(X \wedge E \rightarrow \neg Y) \simeq 0.67$, y $\text{Conf}(Y \wedge E \rightarrow \neg X) \simeq 0.67$, por lo que E estaría en las reglas de excepción de los dos sentidos de la regla doble.

Semánticamente podríamos decir que hemos encontrado un "agente" que influye en los dos sentidos de la regla doble, es decir, su presencia perturba el comportamiento usual de la regla doble y fuerte $X \leftrightarrow Y$. Este caso es muy interesante puesto que la interacción de este "extraño" factor (en el sentido de infrecuente) cambia el comportamiento de la regla doble en los dos sentidos a la vez. Cuando esto suceda, diremos que E define una *excepción doble*.

Para encontrar este tipo de excepciones dobles, se hace indispensable que los grados de confianza de la regla doble en ambos sentidos no estén próximos a 1, ya que cuando esto suceda hay menos probabilidad de que la excepción sea la misma para los dos sentidos de la regla doble.

Para los casos 2 y 3, dependiendo del grado de cumplimiento dado por la medida de interés (confianza o factor de certeza en nuestro caso), obtenemos

X	Y	F	$\cdots$
X	Y	F	$\cdots$
X	Y	F	$\cdots$
X	Y	F	$\cdots$
X	Y	F	$\cdots$
X	Y	F	$\cdots$
X	Y	E	$\cdots$
X	Y'	E	$\cdots$
X	Y'	E	$\cdots$
X'	Y	E	$\cdots$
X'	Y	E	$\cdots$
X'	Y'	E	$\cdots$

Tabla 2.24: Base de datos que tiene una excepción doble.

una descripción más detallada sobre la relación entre dos itemsets asociados de forma doble y fuerte. Este es el caso de nuestro ejemplo, donde la relación entre pájaro y volar está totalmente descrita cuando también conocemos las excepciones asociadas. En otras áreas como por ejemplo en medicina, además de las reglas de excepción, es muy útil obtener las reglas anómalas asociadas a la regla doble.

Reglas Anómalas asociadas a Reglas Dobles

Como sucede en la anterior sección, cuanta más información haya disponible sobre la asociación entre dos conjuntos de items, mejor describiremos su relación. Al igual que sucedía con las reglas de excepción, podemos encontrar reglas anómalas asociadas a cada uno de los dos sentidos de la regla doble, teniendo las siguientes situaciones:

- No hay reglas anómalas asociadas a la regla doble.
- Una de los sentidos de la regla doble tiene asociado alguna regla anómala.
- En ambos sentidos de la regla doble encontramos distintas reglas anómalas, es decir, reglas anómalas con $A \neq A'$.

- Existe un único ítem A que aparece en las reglas anómalas asociadas en ambos sentidos de la regla doble.

Lo más interesante a resaltar aquí es la reformulación de las reglas anómalas usando reglas dobles. Si echamos un vistazo a nuestra propuesta, podría resumirse mediante el cumplimiento de las dos siguientes condiciones:

$X \leftrightarrow Y$	doble fuerte en D
$\neg Y \leftrightarrow A$	doble confidente/fidedigna en D_X

donde la regla $\varphi \leftrightarrow \psi$ es *doble confidente/fidedigna* si tanto $\varphi \rightarrow \psi$ como $\psi \rightarrow \varphi$ son confidentes/fidedignas.

2.4.6. Discusión y Conclusiones

Las reglas de asociación capturan la semántica de la aparición conjunta y frecuente de un conjunto de atributos, pero existen otro tipo de asociaciones que aún siendo infrecuentes también pueden resultar interesantes. Este es el caso de las reglas de excepción y las reglas anómalas. Hemos hecho un amplio recorrido en las diversas propuestas, enfatizando en aquellas que se basan en una formulación usando reglas de asociación, en particular en las que utilizan la terna de reglas $(csr, exc/anom, ref)$.

Por otro lado, hemos introducido el concepto de regla doble que recoge una asociación más íntima entre dos conjuntos de items, que puede ser de gran utilidad cuando dichos itemsets están relacionados en ambos sentidos de la regla. Hemos propuesto un método para conocer de forma más fiable la relación mediante la extracción de las reglas de excepción y anómalas asociadas a la regla doble. Esto puede ser de utilidad e interés en algunos dominios como la medicina, donde sabemos que existen relaciones muy fuertes entre algunos conjuntos determinados de items (enfermedades, síntomas y medicamentos por ejemplo).

Queremos insistir en que la semántica recogida por las reglas de excepción y las anómalas es distinta. Las reglas de excepción requiere la existencia de un ítem (o itemsets) conflictivo al que hemos llamado E, mientras que para extraer reglas anómalas se impone que la mayoría de las excepciones correspondan al mismo consecuente A para poder ser considerada una regla anómala. Para ilustrar estas diferencias, en [Berzal et al., 2004] se presentan varios ejemplos.

Sin embargo, puede ocurrir que la semántica de anomalía o excepción estén mezcladas. En estas situaciones el contexto y la experiencia del usuario o experto en el área son herramientas indispensables para discernir qué tipo de conocimiento se ha extraído. La base de datos mostrada en la Tabla 2.25 es un ejemplo artificial para describir esta posibilidad bajo ciertas hipótesis sobre los umbrales $minconf$ o $minFC$ impuestos.

X	Y	Z_1	$\cdots$
X	Y	Z_1	$\cdots$
X	Y	Z_2	$\cdots$
X	Y	Z_2	$\cdots$
X	Y	Z_3	$\cdots$
X	Y	Z_3	$\cdots$
X	Y	Z_4	$\cdots$
X	Y	Z	$\cdots$
X	Y'	Z	$\cdots$
X	Y'	Z	$\cdots$
X'	Y	Z	$\cdots$
X'	Y	Z	$\cdots$
X'	Y	Z	$\cdots$
X'	Y	Z	$\cdots$

Tabla 2.25: ¿Z está asociada a una regla de excepción o a una regla anómala?

De esta tabla usando los umbrales 0.3 para el soporte y 0.6 para la confianza, obtenemos que $\text{Sop}(X \rightarrow Y) = 0.571$ y $\text{Conf}(X \rightarrow Y) = 0.8$ por lo que $X \rightarrow Y$ es una regla common sense (regla fuerte). El ítem Z es una excepción usando por ejemplo el enfoque de Hussain et al. [Hussain et al., 2000] puesto que $\text{Conf}(X \wedge Z \rightarrow \neg Y) = 0.667$ y su condición para la regla de referencia también se cumple ya que $\text{Sop}(Z \rightarrow Y) = 0.357$ y $\text{Conf}(Z \rightarrow Y) = 0.714$ indican que $Z \rightarrow Y$ es fuerte. Pero observemos que $X \rightsquigarrow Z \,|\neg Y$ satisface las condiciones para ser una regla anómala porque $\text{Conf}(X \wedge \neg Y \rightarrow Z) = 1$ y $\text{Conf}(X \wedge Y \rightarrow \neg Z) = 0.875$ si usamos la propuesta de Berzal et al. y $\text{Conf}(X \wedge Z \rightarrow \neg Y) = 0.667$ con nuestra propuesta. Sin embargo, usando el factor de certeza, $FC(X \wedge Z \rightarrow X \wedge \neg Y) = 0.571$

el valor es menor y por tanto descartaríamos a ésta última regla como anómala.

Cuando haya situaciones en las que no sepamos distinguir si tenemos una regla de excepción o una regla anómala, la forma más razonable de proceder es echando una mirada al contexto y después decidir la semántica que mejor se ajuste según el atributo o ítem Z en cuestión. Cuando queramos utilizar una condición objetiva para discernir entre una u otra, podemos elegir de acuerdo al mayor valor obtenido en la medida de interés, o bien cambiar esta por una medida más fuerte como proponemos nosotros usando el factor de certeza.

Aunque la semántica recogida por ambas reglas (de excepción y anómala) es distinta, ambas miden un comportamiento 'extraño', novedoso e infrecuente alrededor de la regla common sense, y esto puede ser de utilidad en diferentes escenarios.

Concluimos la sección destacando los puntos más interesantes de nuestras propuestas.

- Puesto que $\text{Conf}(X \wedge A \rightarrow \neg Y) \leq \text{Conf}(X \wedge Y \rightarrow \neg A)$ se cumple bajo una condición que es cierta en general (ver proposición 2.6), si evaluamos la validez de las reglas anómalas con nuestra propuesta, las reglas obtenidas también cumplen las condiciones dadas en [Berzal et al., 2004]. Por tanto, nuestra propuesta obtiene reglas anómalas más fiables siendo el número total de reglas obtenidas mucho menor. Además usando el factor de certeza, disminuiremos aún más su número puesto que es una medida más fuerte que la confianza.

- Para distinguir cuándo un ítem contribuye en una regla de excepción o en una anómala, podemos utilizar el factor de certeza que es más preciso y más restrictivo que la confianza.

2.5. Resumen

El almacenamiento de información en bases de datos de distinta índole es muy común en nuestros días. Por ello es necesario tener a nuestro alcance distintos tipos de herramientas que se adapten tanto al tipo de dato almacenado como al conocimiento que el usuario pretende extraer de dicho sistema de información.

Este capítulo ha presentado nuevas herramientas para obtener diversas clases de conocimiento adaptándose cuando es preciso al tipo especial de almacenamiento como han sido las bases de datos formadas por bolsas.

Además se ha desarrollado el modelo formal como marco unificado para trabajar con cualquier tipo de regla de asociación, adaptándolo cuando ha sido necesario para reglas con más complejidad como las reglas de excepción y las reglas anómalas. El modelo nos proporciona tres importantes propiedades: un modelo en el que representar de forma sencilla la información, un marco para el estudio de las distintas medidas de evaluación para reglas de asociación y una notación unificada que nos va a servir en el capítulo siguiente para desarrollar un procedimiento único para extraer distintos tipos de reglas de asociación de una base de datos dependiendo solamente de los cuantificadores-4ft y los umbrales asociados elegidos por el usuario.

Concretamente, se han obtenido resultados muy interesantes que relacionan las propiedades de un cuantificador-4ft con los principios de bondad de una medida de interés. En particular, hemos probado que el factor de certeza visto como cuantificador-4ft cumple los distintos principios propuestos en la literatura por Piatetsky-Shapiro junto con los dos nuevos que hemos propuesto y justificado en la sección 2.1.5.

También hemos propuesto una generalización del modelo para reglas difusas. Dicha propuesta nos permite obtener nuevas medidas de interés para extraer este tipo de reglas usando el modelo de representación mediante niveles de restricción en vez de la teoría de conjuntos difusos. Esto nos permite manejar negaciones de items respetando las propiedades deseables que teníamos cuando extraemos reglas de asociación crisp.

Siguiendo el uso de medidas de interés para reglas difusas, hemos presentado varios procedimientos para capturar información de forma cuantitativa cuando la base de datos está formada por bolsas. Puesto que las bolsas son una generalización del concepto de transacción y almacenan información relacionada con la cantidad de los items, el tipo de conocimiento obtenido es más rico y ofrece no sólo relaciones entre los items, si no también entre sus cantidades, utilizando para ello reglas

difusas y dependencias graduales adaptadas a este tipo de transacción.

Por último, pero no por ello menos interesante, hemos tratado con un tipo nuevo de información, que es aquella que es infrecuente y que ofrece alguna peculiaridad que puede ser de interés para el usuario. El marco de trabajo que hemos utilizado es trabajar con un conjunto de reglas que cumplen una serie de propiedades y que se ajustan a una semántica de interés para el usuario según el dominio donde se aplique. El objetivo es obtener reglas de excepción o reglas anómalas que sean fiables. Además hemos presentado un proceso por el que podemos estudiar con mayor profundidad el tipo de asociación o relación entre dos itemsets cuando éstos presentan una asociación fuerte en ambos sentidos de la regla de asociación mediante el uso conjunto de reglas dobles con sus reglas de excepción y anómalas asociadas.

Capítulo 3

Experimentos y Resultados

En este capítulo presentamos un algoritmo basado en el modelo formal para obtener reglas de asociación a partir de los cuantificadores-4ft. En particular el algoritmo está descrito para obtener las reglas de excepción y las reglas anómalas asociadas a una regla de asociación sencilla o bien a una regla doble.

El algoritmo sirve como apoyo práctico para obtener un estudio cualitativo y cuantitativo de nuestras propuestas así como una comparación con anteriores trabajos. Aun habiendo estudiado teóricamente algunas de las propiedades de nuestras propuestas, se hace necesario una implementación práctica puesto que el objetivo de las herramientas desarrolladas tienen como fin ser usadas en bases de datos reales para obtener el máximo conocimiento extraíble de ellas.

Uno de los objetivos que queremos conseguir con la implementación que presentaremos a continuación es comprobar que el modelo ofrece un marco unificado para representar los diversos tipos de reglas de asociación utilizando los cuantificadores-4ft que convengan. Otro de los objetivos es mostrar que la extracción tanto de reglas de excepción como de reglas anómalas es plausible en la práctica para bases de datos reales y de gran tamaño con consumos de tiempo y memoria admisibles.

El capítulo sigue con un vistazo a los principales algoritmos de extracción de reglas de asociación, deteniéndose en una propuesta que utiliza una representación binaria de los items [Louie and Lin, 2000b] que nos será de utilidad para desarrollar nuestra propuesta en la sección siguiente. Seguiremos con la descripción de nuestro procedimiento para extraer reglas de excepción, anómalas y dobles, y terminaremos

con una comparativa de los resultados obtenidos por nuestra propuesta y las de otros autores.

3.1. Algoritmos de Extracción de Reglas de Asociación

El principal reto cuando extraemos reglas de asociación es el inmenso número de reglas que pueden considerarse. De hecho, el número de reglas obtenidas crece exponencialmente con el número de items. Por ello se hace indispensable restringir el número de reglas mediante el uso de umbrales mínimos para las medidas de interés que se utilicen en el proceso de minería. Esta restricción nos permite dividir el BARP (Binary Association Rule Problem) en dos subproblemas como ya mencionamos en el primer capítulo y que se conoce como Algoritmo Apriori Básico.

Algoritmo 3.1 : Apriori Básico

Entrada: I, D, $minsop$, $minconf$

Salida: Conjunto de reglas de asociación con soporte y confianza $\geq minsop$ y $minconf$.

Algoritmo:

1. Generar todos los itemsets frecuentes, i.e. con soporte $\geq minsop$.
2. Dado un itemsets frecuente $X = \{x_1, \ldots, x_k\}$ con $k \geq 2$, generar todas las reglas de la forma $X \backslash \{x_j\} \rightarrow \{x_j\}$, siendo el soporte de dicha regla el soporte de X; y la confianza, el cociente de este soporte y el soporte de $X \backslash \{x_j\}$.

Para determinar la confianza en este caso, es suficiente con conocer el soporte de los subconjuntos de X. El conocimiento de los valores de soporte está asegurado por la propiedad de *clausura descendente* que cumple el soporte:

> *"Todos los subconjuntos de un itemsets frecuente deben ser también frecuentes"* [Agrawal and Srikant, 1994].

El problema en nuestro caso difiere un poco al planteamiento inicial de la extracción de reglas aunque se sirve de éste puesto que las reglas de tipo common sense son reglas fuertes que pueden extraerse mediante cualquier tipo de algoritmo de extracción de reglas de asociación.

Los distintos algoritmos propuestos en la literatura pueden clasificarse en dos grandes clases según la estrategia principal que sigan [Hipp et al., 2000]:

- BFS (Breadth-first search): Se determina el soporte de todos los $(k-1)$-itemsets antes de medir el soporte de los k-itemsets.

- DFS (Depth-first search): El soporte se calcula recursivamente descendiendo en una estructura de tipo árbol.

A su vez cada uno puede dividirse en dos subtipos según utilicen el conteo de ocurrencias o la intersección de los k-itemsets como podemos observar en la Figura 3.1. A continuación presentamos con más detalle los más representativos de cada tipo.

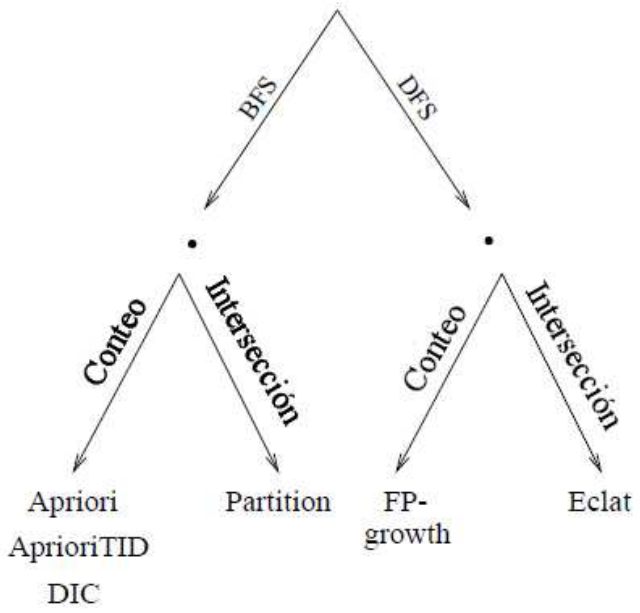

Figura 3.1: Diagrama con los tipos de algoritmos de extracción de reglas de asociación.

Apriori [Agrawal and Srikant, 1994] cuenta todos los candidatos con cardinalidad k en un solo pase de la base de datos. La parte crítica es buscar dichos candidatos en cada una de las transacciones. Para ello en [Agrawal and Srikant, 1994] se introduce la llamada estructura *hashtree* utilizando una serie de prefijos para la búsqueda de dichos candidatos en las transacciones. **AprioriTID** [Agrawal and Srikant, 1994] es una extensión del algoritmo básico Apriori, pero utilizando una representación interna de cada transacción mediante los candidatos que esta contiene. También podemos encontrar más variantes de estos algoritmos como el **AprioriHybrid** [Agrawal and Srikant, 1994], combinando los dos

anteriores, o **SETM** [Houtsma and Swami, 1993] que es del mismo tipo pero implementado directamente en SQL.

Otra variante del tipo Apriori es **DIC** (Dynamic Itemsets Counting) [Brin et al., 1997] que suaviza la estricta separación entre los procesos de conteo y de generación de candidatos mediante el llamado *prefix-tree* que difiere del hashtree en que los candidatos pueden encontrarse en cualquier nodo del árbol en vez de sólo en las hojas.

El algoritmo **Partition** [Savasere et al., 1995] es del tipo Apriori pero en él se utiliza la intersección de conjuntos para determinar los soportes de los itemsets.

El principal problema de los algoritmos de tipo BFS es el elevado número de pases de la base de datos ya que para cada candidato de tamaño k se escanea una vez la base de datos. En [Pei et al., 2004] se presenta un nuevo tipo de algoritmo llamado **FP-growth** del tipo DFS. Este introduce un nuevo tipo de árbol *FP-tree* que se construye mediante el conteo de ocurrencias. En el FP-tree se condensa la información de las transacciones y a partir de él se puede calcular el soporte de todos los itemsets frecuentes.

En [Zaki et al., 1997] se introduce el algoritmo **ECLAT** que combina DFS con intersecciones de tidlists[1] empleando un tipo de optimización llamado "intersecciones rápidas". Este algoritmo también posee variantes como **MaxEclat** para extraer aquellos itemsets frecuentes que sean maximales [Zaki et al., 1997].

Para una descripción más detallada de estos algoritmos podemos consultar [Goethals, 2003] y [Hipp et al., 2000] además de las citadas referencias. También podemos encontrar un estudio comparativo de los anteriores algoritmos.

Además de los algoritmos que hemos mencionado, existen otras propuestas que difieren bastante de la dinámica común de todos ellos. Destacaremos en particular la propuesta dada en [Louie and Lin, 2000b] usando computación mediante bits. En las secciones siguientes describiremos las principales características de este algoritmo que luego adaptaremos para la extracción de reglas de excepción y anómalas ya que se adapta muy bien al modelo formal que hemos ido desarrollando en los anteriores capítulos.

[1] Tid es un identificador único para una transacción. Una tidlist (asociada a un ítem) es un conjunto de identificadores que corresponden a las transacciones que contienen a dicho ítem.

3.1.1. Búsqueda de Reglas de Asociación usando Computación mediante Bits

El estudio de los modelos orientados a la máquina llevado a cabo por varios autores en [Louie and Lin, 2000a], [Lin, 2000], [Louie and Lin, 2000b] ofrece una nueva perspectiva en el campo de la Minería de Datos.

La idea principal de este modelo recae en el concepto de gránulo o clase de equivalencia. Un par $(atributo, valor)$ puede verse como una colección de gránulos que poseen el mismo tipo de propiedad. Estos gránulos se representan mediante conjuntos de bits y de este modo, el problema de encontrar reglas de asociación se transforma en un conjunto de operaciones booleanas entre los conjuntos de bits, que de ahora en adelante llamaremos *bitsets*.

Sea U un conjunto que denota un universo formado por entidades u objetos. En bases de datos relacionales, U será el conjunto de entidades que están representadas por la relación. Los valores de los atributos corresponden a las propiedades de las entidades y se representan por los denominados *conceptos elementales*. C será el conjunto de todos los conceptos elementales.

Un atributo A induce una relación de equivalencia en U como sigue: dos tuplas (entidades) son equivalentes (con la relación inducida por A) si y sólo si todos sus valores coinciden con los de A. Esta relación de equivalencia crea una partición de U en clases de equivalencias disjuntas a las que llaman *gránulos* en [Louie and Lin, 2000b]. Un gránulo puede representarse por una lista de identificadores de tuplas o mediante una representación binaria. En forma binaria, el valor 1 ó 0 en una posición determinada indicará si la tupla satisface un atributo particular o no respectivamente.

Un ejemplo nos servirá para aclarar las ideas. Consideremos una relación consistente en dos columnas, ID del vehículo y el tipo de vehículo como en la Tabla 3.1.

Los valores de la primera columna de la tabla identifican de forma única las tuplas (entidades). Esta característica donde todos los valores de los atributos son únicos no es de utilidad en minería de datos puesto que los valores tienen correspondencia 1 a 1 a ellos mismos, pero en este modelo se utilizará para hacer referencia a las tuplas. La columna tipo de vehículo no es como la anterior y contiene mucha información de utilidad para los algoritmos de extracción de reglas de asociación. Esta columna tiene cinco tipos distintos de vehículos, que da lugar al conjunto cociente {turismo, deportivo, monovolumen, coupé, todoterreno} donde

ID vehículo	Tipo de vehículo
id_1	turismo
id_2	deportivo
id_3	turismo
id_4	monovolumen
id_5	coupé
id_6	monovolumen
id_7	deportivo
id_8	turismo
id_9	turismo
id_{10}	deportivo
id_{11}	todoterreno
id_{12}	turismo

Tabla 3.1: Base de datos relacional con dos columnas.

cada elemento de este conjunto representa a un gránulo o clase de equivalencia. En la Tabla 3.2 podemos ver una partición del conjunto en dos formas distintas: mediante listas y con una representación binaria.

Representantes	Repres. por Listas	Repres. Binaria
turismo	$\{id_1, id_3, id_8, id_9, id_{12}\}$	101000011001
deportivo	$\{id_2, id_7, id_{10}\}$	010000100100
monovolumen	$\{id_4, id_6\}$	000101000000
coupé	$\{id_5\}$	000010000000
todoterreno	$\{id_{11}\}$	000000000010

Tabla 3.2: Clases de Equivalencia inducidas por la Tabla 3.1 con dos tipos de representaciones.

Observemos que cada elemento en el conjunto cociente es un subconjunto del universo U. El número de elementos en el conjunto cociente es, por lo general, mucho menor que el número de filas de la base de datos. Por lo tanto, la computación granular o mediante gránulos reduce el problema a un dominio de menor tamaño, a saber, una tabla con múltiples columnas se reduce a una familia de particiones.

Este tipo de reducción puede usarse para la extracción de reglas de asociación como sigue. En términos de gránulos, (X, Y) es una regla de asociación si y sólo si:

1. el porcentaje de bits con valor 1 de la intersección $X \cap Y$ es mayor o igual que el porcentaje impuesto para el mínimo soporte, y

2. el porcentaje de bits con valor 1 de $X \cap Y$ entre el porcentaje de bits con valor 1 de X es mayor o igual que el porcentaje impuesto para la mínima confianza.

La intersección $X \cap Y$ puede ser generalizada a un número finito de intersecciones de gránulos: $X_1 \cap \ldots \cap X_i \cap Y_1 \cap \ldots \cap Y_j$. La intersección entre gránulos puede identificarse con el operador booleano AND usando la representación binaria. Este operador es menos costoso computacionalmente que la intersección, por lo que nos reducirá el tiempo de ejecución para calcular el soporte de los

itemsets y a partir de él podremos calcular la confianza. Veamos el proceso mediante un ejemplo.

Ejemplo 3.1. Siguiendo con la relación entre vehículos mostrada en la Tabla 3.1, consideremos la extensión de dicha relación mediante dos columnas más dadas en la Tabla 3.3.

ID vehículo	Tipo de vehículo	Color	Precio
id_1	turismo	rojo	moderado
id_2	deportivo	azul	caro
id_3	turismo	verde	caro
id_4	monovolumen	blanco	moderado
id_5	coupé	rojo	barato
id_6	monovolumen	rojo	caro
id_7	deportivo	negro	moderado
id_8	turismo	azul	caro
id_9	turismo	verde	caro
id_{10}	deportivo	negro	moderado
id_{11}	todoterreno	verde	moderado
id_{12}	turismo	azul	caro

Tabla 3.3: Relación entre vehículo y sus características

Las clases de equivalencia (gránulos) que obtenemos de las dos nuevas columnas son {rojo, azul, verde, blanco, negro} y {barato, moderado, caro} con representaciones asociadas en la Tabla 3.4.

Para obtener combinaciones entre dos clases de equivalencia provenientes de dos columnas (atributos) distintas usando intersecciones, se utilizará el operador booleano AND cuyo funcionamiento puede verse de forma sencilla con la representación binaria. Otros operadores como OR y la negación NEG también pueden definirse de forma sencilla. En la Tabla 3.5 se muestran las tablas para definir dichos operadores.

De esta forma, la intersección de gránulos se efectuará con el operador AND, la unión mediante el operador OR y para el complementario usaremos la negación

Representantes	Repres. por Listas	Repres. Binaria
rojo	$\{id_1, id_5, id_6\}$	100011000000
azul	$\{id_2, id_8, id_{12}\}$	010000010001
verde	$\{id_3, id_9, id_{11}\}$	001000001010
blanco	$\{id_4\}$	000100000000
negro	$\{id_7, id_{10}\}$	000000100100
barato	$\{id_5\}$	000010000000
moderado	$\{id_1, id_4, id_7, id_{10}, id_{11}\}$	100100100110
caro	$\{id_2, id_3, id_6, id_8, id_9, id_{12}\}$	011001011001

Tabla 3.4: Clases de Equivalencia inducidas por las dos últimas columnas de la Tabla 3.3.

X	Y	$X\ AND\ Y$	$X\ OR\ Y$	$NEG\ X$
0	0	0	0	1
1	0	0	1	0
0	1	0	1	1
1	1	1	1	0

Tabla 3.5: Operaciones AND, OR y NEG para bits.

Combinación	Repres. Binaria	Resultado	Sop
deportivo ∩ caro	010000100100 AND 011001011001	010000000000	1/12
monovol. ∩ caro	000101000000 AND 011001011001	000001000000	1/12
coupé ∩ caro	000010000000 AND 011001011001	000000000000	0
todoterreno ∩ caro	000000000010 AND 011001011001	000000000000	0
turismo ∩ caro	101000011001 AND 011001011001	001000011001	4/12

Tabla 3.6: Intersección entre los gránulos del atributo *tipo de vehículo* y el gránulo *caro*.

NEG. Aunque para el cálculo del soporte y la confianza nos bastará utilizar el operador AND, para otras medidas será necesario el uso del resto de operadores. En el ejemplo que estábamos analizando, podemos ver que la intersección de los gránulos procedentes del atributo *tipo de vehículo* con el gránulo *caro* daría lugar a cinco intersecciones cuyo soporte podemos ver en la Tabla 3.6.

♦

Mediante este tipo de representación denominada *computación mediante gránulos* o *granular computing* es fácil obtener un sencillo algoritmo, que sigue en esencia los pasos del algoritmo Apriori, para descubrir reglas de asociación en una base de datos pero las operaciones se llevan a cabo con las representaciones binarias asociadas. Si utilizamos el marco soporte/confianza, utilizaremos el operador AND y el $Cardinal$ de un gránulo que no es más que contar el número de unos de una representación binaria para obtener el soporte y partir de éste la confianza. Con estos dos operadores, y sabiendo el número total de transacciones, es decir, el número de objetos del universo U, el cómputo de los valores para el soporte y la confianza son muy sencillos:

$$\text{sop}(X) = \frac{Cardinal\{Bin(X)\}}{|U|} \tag{3.1}$$

$$\text{Sop}(X \rightarrow Y) = \frac{Cardinal\{Bin(X)\ AND\ Bin(Y)\}}{|U|} \tag{3.2}$$

$$\text{Conf}(X \rightarrow Y) = \frac{\text{Sop}(X \rightarrow Y)}{\text{sop}(X)} \tag{3.3}$$

donde $Bin(\cdot)$ denota la representación binaria de un gránulo o clase de equivalencia, $Cardinal\{\cdot\}$ es el número de unos de la correspondiente representación

binaria y $|U|$ es el número de objetos del universo U que coincide con el número de transacciones de una base de datos.

Hay dos propiedades muy interesantes a tener en cuenta de estas representaciones binarias:

1. Las representaciones binarias asociadas a un mismo atributo (columna) forman una partición de la base de datos (del conjunto de transacciones) siempre y cuando no falten valores en la base de datos. Cuando esto ocurra, podemos definir un nuevo ítem del tipo $\langle atributo, ? \rangle$ para obtener una partición.

2. La conjunción mediante AND de dos clases de equivalencia asociadas a la misma columna o de dos representaciones binarias del mismo atributo da la representación binaria nula: $\overbrace{00\ldots00}^{n}$, donde n es el número de transacciones en la base de datos.

En [Louie and Lin, 2000b] y [Rauch and Šimunek, 2005] podemos encontrar dos propuestas distintas que utilizan granular computing para obtener reglas de asociación en una base de datos. La primera propuesta [Louie and Lin, 2000b] puede seguir dos estrategias para realizar el cálculo de las intersecciones de los gránulos:

1. Se calcula la intersección completa entre dos gránulos

2. Se va calculando la intersección entre dos gránulos hasta que se alcanza la condición del mínimo soporte.

La segunda opción acelera el proceso ya que una vez obtenido que el itemsets es frecuente, ya podemos utilizarlo para obtener el conjunto de k-itemsets frecuentes. La principal desventaja es que para obtener el nivel siguiente, los $(k+1)-$itemsets, es necesario conocer la representación binaria del k-itemsets predecesor.

La segunda propuesta [Rauch and Šimunek, 2005] utiliza la representación binaria mediante lo que los autores llaman *cadenas de bits*. Cada cadena de bits representa a un itemsets del tipo $\langle atributo, valor \rangle$ o $\langle atributo, intervalo \rangle$ y se corresponde con los gránulos que se utilizan en [Louie and Lin, 2000b]. Esta propuesta difiere en la anterior en que también utiliza la negación NEG para obtener reglas de asociación que involucran tanto items como sus negaciones.

El procedimiento que presentan los autores en [Rauch and Šimunek, 2005] denominado como **4ft-Miner**, necesita además que se le especifiquen con

antelación cuáles son el conjunto de antecedentes y consecuentes que se quieren considerar en el proceso de extracción de reglas. Con dichos conjuntos y la base de datos asociada, calculan las tablas-4ft para cada par de itemsets ($X \in \{antecedentes\}, Y \in \{consecuentes\}$) y después calculan el valor de los cuantificadores de cada asociación del tipo $X \approx Y$ extrayendo sólo aquellas que superen los umbrales pre-establecidos.

La siguiente sección mostrará un nuevo procedimiento para obtener reglas de asociación así como las reglas de excepción y anómalas asociadas mediante el uso de la representación binaria presentada aquí.

3.2. Extracción de Reglas de Asociación a través del Modelo Lógico.

El modelo formal desarrollado en el Capítulo 2 para la representación y evaluación de reglas de asociación nos permite obtener distintos tipos de asociaciones entre grupos de items incluyendo además de la conjunción de items, su disyunción y negación. En particular, las reglas de excepción y las reglas anómalas hacen uso de la negación de items para su formulación.

Cuando la negación entra en juego en el descubrimiento de reglas de asociación el coste computacional crece puesto que el número de items que interviene es el doble y además el soporte de la negación de dichos items generalmente supera el umbral del mínimo soporte. Este fenómeno es debido a que si $\text{sop}(X) < minsop$ entonces $\text{sop}(\neg X) = 1 - \text{sop}(X) > minsop$ si $minsop \leq 0.5$. Sin embargo el uso de la negación está debidamente justificado en las definiciones de reglas anómalas y de excepción, consiguiendo que las reglas obtenidas tengan un contenido semántico útil para el usuario.

El problema que se presenta en estos dos tipos de reglas, de excepción y anómalas, es que la condición que deben cumplir no involucra al soporte si no solamente a la medida de interés que utilicemos: confianza o factor de certeza en nuestro caso. Esto hace que los criterios de poda relacionados con el soporte no puedan utilizarse y por tanto, los algoritmos de búsqueda suelen ser bastante lentos computacionalmente hablando.

El algoritmo que presentaremos a continuación trata de mejorar los tiempos de computación gracias a la computación granular que hemos presentado en la sección 3.1.1 así como al uso del factor de certeza que al reducir sustancialmente el número de reglas de asociación, también se reduce el coste de buscar para cada regla common sense las reglas anómalas y de excepción asociadas.

3.2.1. Algoritmo e Implementación

Nuestra propuesta es capaz de extraer conjuntamente un conjunto de reglas fuertes con sus reglas anómalas y de excepción asociadas. Explicaremos el funcionamiento de nuestro algoritmo y su implementación para reglas de excepción y después indicaremos los cambios necesarios para obtener reglas anómalas.

En el proceso de extracción de las reglas de excepción consideraremos que $E \in I$ posea un sólo ítem, que por supuesto no tiene por qué ser frecuente. La

principal novedad de nuestro algoritmo es el uso de BitSets para representar a los items y a los itemsets que se van formando durante el proceso. A diferencia de los algoritmos presentados en [Louie and Lin, 2000b] y [Rauch and Šimunek, 2005] que utilizan una representación mediante una cadena de bits, nosotros hacemos uso de la clase `java.util.BitSet` ya implementada en el lenguaje java llamada BitSet que contiene el objeto `BitSet` y algunas operaciones de interés: AND, XOR, OR, Cardinality, etc. pero no posee la negación de un BitSet por lo que tenemos que definirla nosotros. Hemos observado que con nuestra definición de negación el cómputo de la negación del bitset ralentiza el proceso de minería por lo que hemos optado por obtener los cardinales necesarios utilizando la tabla-4ft y las relaciones entre las frecuencias.

El objeto BitSet almacena un conjunto de bits (ceros y unos) de forma ordenada. En la implementación de nuestro algoritmo, para cada ítem de la forma $\langle atributo, valor\rangle$ o $\langle atributo, intervalo\rangle$ creamos un BitSet de tamaño igual al número de transacciones de la base de datos que contendrá en la posición i-ésima si dicho ítem (o itemsets si tenemos un conjunto de items) se satisface en la transacción i-ésima (1) o no (0).

El proceso para obtener reglas de excepción al que hemos llamado **ERSA** (Exception Rule Search Algorithm) se describe en el Algoritmo 3.2.1 y está compuesto de tres pasos bien diferenciados:

1. Preprocesamiento de la base de datos para transformarla en un conjunto de clases de equivalencia.

2. Extracción del conjunto de reglas common sense (csr).

3. Para cada csr obtenida buscamos sus reglas de excepción.

El primer paso se puede llevar a cabo de distintas formas. En nuestro caso, hemos utilizado un fichero de texto con la información necesaria sobre los items de la base de datos. Debido a la heterogeneidad de los datos que puede poseer una base de datos, numéricos, categóricos, continuos, etc. esta es una buena solución para tratar los distintos tipos de items obteniendo dos tipos principalmente: $\langle atributo, valor/categoría\rangle$ y $\langle atributo, intervalo\rangle$.

Los pasos 1.1 y 1.3 del Algoritmo 3.2.1 son ejecutados conjuntamente para leer una sola vez la base de datos. A partir de ella se crea un vector de BitSets con dimensión igual al número de transacciones y donde cada elemento del vector hace referencia a un ítem particular. Cada BitSet contendrá el valor uno o cero

Algoritmo 3.2 : Búsqueda de Reglas de Excepción (ERSA)

Entrada: Base de Datos Transaccional, $minsop$, $minconf$ o $minFC$

Salida: Conjunto de reglas de asociación con sus reglas de excepción asociadas.

1. **Preprocesamiento de la Base de Datos**
   1.1 Transformación de la base de datos en una base de datos booleana.
   1.2 Almacenamiento de la base de datos en un vector de BitSets.
2. **Proceso de Minería.**
   2.1 **Extracción de las Reglas Common Sense (*csr*)**
   Búsqueda del conjunto de candidatos (itemsets frecuentes)
   para extraer las *csr*.
   Almacenamiento de los índices y el soporte asociado a
   los itemsets frecuentes del vector de Bitsets.
   Extracción de las *csr* que superen los umbrales
   de $minsop$ y $minconf/minFC$.
   2.2.1 **Extracción de las Reglas de Excepción**
   Para cada *csr*, $X \rightarrow Y$, buscamos las posibles excepciones:
   Para cada $E \in I$ (excepto aquellos cuyo atributo esté en la *csr*)
   Calculamos $X \wedge E \wedge \neg Y$ y su soporte
   Calculamos $X \wedge E$ y su soporte
   Si utilizamos la confianza:
   Calculamos $\text{Conf}(X \wedge E \rightarrow \neg Y)$ **si** es $\geq minconf$
   entonces hemos encontrado una regla de excepción
   Si utilizamos el factor de certeza:
   Calculamos $\text{sop}_X(\neg Y)$
   Calculamos $FC_X(E \rightarrow \neg Y)$ **si** es $\geq minFC$
   entonces hemos encontrado una regla de excepción

en la i-ésima posición si el ítem se satisface o no respectivamente en la i-ésima transacción.

El paso 2.1 obtiene las reglas common sense mediante la conjunción de dos o más itemsets frecuentes de los que se almacena en un vector independiente su índice en el vector de BitSets y su soporte. Para obtener las reglas common sense será necesario que se cumplan los umbrales predefinidos para el soporte y la confianza (o el factor de certeza). En nuestro proceso se tendrán en cuenta ambas direcciones de las reglas siempre que cumplan la condición anterior, para así obtener las reglas que sean dobles.

Una vez obtenida una regla common sense se empieza el proceso para obtener las reglas de excepción. El paso 2.2.1 describe este proceso. Para ello se utilizan los BitSets asociados al antecedente y al consecuente de la regla common sense, llamémosles X e Y respectivamente. El punto crucial del paso 2.2.1 es el de calcular la confianza/factor de certeza de la regla de excepción para cada ítem de I exceptuando aquellos cuyo atributo coincide con alguno de los de la regla common sense. Es decir, debemos calcular $\mathrm{Conf}(X \wedge E \rightarrow \neg Y)$ o bien $FC(X \wedge E \rightarrow \neg Y)$ donde $E \in I$ es un ítem. Este proceso es muy costoso y depende fuertemente del número de items de I puesto que hay que calcular la conjunción $X \wedge E$ para cada $E \in I$ exceptuando los que tienen atributo común con la csr.

Una de las mayores ventajas de usar BitSets es que las operaciones de conjunción y cardinal son muy rápidas. El problema es el cálculo de la negación $\neg Y$ que es mucho más lenta que la conjunción. Para solventar este inconveniente, utilizaremos lo que sabemos del modelo formal, en particular de la tabla $4ft(E, Y, D_X) = \mathcal{M}_X^E = \langle e, f, g, h \rangle$ (ver la Sección 2.4.3). De esta forma en vez de hacer la negación de Y, calcularemos las dos frecuencias siguientes $cardinal(X \wedge E \wedge \neg Y)$ y $cardinal(X \wedge E)$, que corresponden con f y $e + f$ de la tabla $4ft(E, Y, D_X)$ (que aparecían para calcular el E-cuantificador $\Rightarrow_I^E$, ecuación (2.36)), como sigue:

$$\begin{aligned} f + e &= cardinal(X \wedge E), \\ f &= cardinal(X \wedge E \wedge \neg Y) = cardinal(X \wedge E) - cardinal(X \wedge E \wedge Y). \end{aligned} \tag{3.4}$$

Para el factor de certeza, necesitamos calcular las cuatro frecuencias e, f, g, h de la tabla-$4ft$ $\mathcal{M}_Z^E$, que utilizando sólo conjunciones y cardinales, primero debemos calcular $e + f$ y e como en la ecuación (3.4), y las sumas

$$\begin{aligned} e + g &= cardinal(X \wedge Y), \text{ y} \\ a + b &= e + f + g + h = cardinal(X) \end{aligned} \tag{3.5}$$

ya están calculadas por la common sense rule. Con estas cantidades es fácil obtener las cuatro frecuencias: $f = (e+f)-e$, $g = (e+g)-e$, $h = (e+f+g+h)-e-f-g$. Por lo tanto, para el cálculo del factor de certeza no hay que calcular ninguna conjunción más que para la confianza entre BitSets y simplemente se realizan algunas operaciones aritméticas más.

Como ya hemos comentado anteriormente, **ERSA** está pensado para extraer las reglas dobles ya que para cada posible regla common sense basta comprobar ambas direcciones de la regla. Esta peculiaridad del algoritmo hace que las reglas extraídas puedan contener más de un ítem en el consecuente de la regla. Debido a este fenómeno, para que las *csr* obtenidas no sean demasiados complejas, fijaremos el número máximo de items en un itemsets frecuente a un número entre 1 y 3.

Para descubrir las reglas anómalas, **ERSA** puede modificarse pasando a llamarse **ARSA** (Anomalous Rule Search Algorithm) cambiando el paso 2.2.1 para obtener las reglas anómalas. El proceso es muy parecido. Tomamos de nuevo items sencillos, $A \in I$ (aquí no exigimos que no tenga atributo común con los items de la *csr*), y después calculamos el valor de los cuantificadores tanto para la regla anómala como para la regla de referencia. El cómputo de dichos valores se efectúa de forma similar al caso de las reglas de excepción mediante conjunciones y cardinales.

Aunque nuestra propuesta utiliza predominantemente el factor de certeza como cuantificador, el algoritmo puede modificarse para cualquier otro tipo de cuantificador dependiendo de las necesidades del usuario y del tipo de reglas que pretenda extraer.

3.2.2. Complejidad de ERSA y ARSA

Los algoritmos **ARSA** y **ERSA** integran la extracción de las reglas anómalas y de excepción durante el proceso de extracción de las reglas common sense. Para cada *csr* encontrada, llamamos al proceso 2.2.1 para extraer la reglas anómalas o de excepción. La complejidad de nuestro algoritmo dependerá en primera instancia del número de transacciones n y del número de items i obteniendo $O(n2^i)$ como en un algoritmo del tipo **Apriori**, pero además dependerá también en segunda instancia del número de reglas *csr* extraídas (r). Por lo tanto en nuestro caso la complejidad teórica es del orden $O(nri2^i)$. Aunque teóricamente el tiempo es muy elevado, en los experimentos llevados a cabo con varios conjuntos de datos reales veremos que el algoritmo conlleva tiempos bastante razonables. De hecho, los dos factores más influyentes en el tiempo de ejecución son el número de reglas

extraídas y el número de items como comprobaremos más adelante.

Tanto **ARSA** como **ERSA** tienen un alto consumo de memoria puesto que el vector de BitSets asociado a la base de datos se almacena en memoria. No obstante para bases de datos standard no hay ningún problema. Por ejemplo para la base de datos `Barbora` (http://lispminer.vse.cz/download) que se utilizó en el congreso PKDD99 en Praga [Rauch and Šimunek, 2005] consistente en 6181 transacciones y 12 atributos (33 items), el vector de BitSets ocupa 107 kb, y para 61810 transacciones es 1.04 MB.

Nombre	Transacciones	Atributos	Items
`Barbora`	6181	7	35
`Mushroom`	8124	23	127
`Breast Cancer`	286	10	53
`Car`	1728	7	25
`Contraceptive`	1473	10	35
`Hepatitis`	155	20	75
`Nursery`	12960	8	32
`Pima Diabetes`	768	9	44
`Post Operative`	90	9	27
`Vote`	435	17	34
`Wisconsin Breast Cancer`	699	10	47
`Abalone`	4177	9	46

Tabla 3.7: Descripción de las Bases de Datos utilizadas en los experimentos.

Para ilustrar el tiempo de ejecución de nuestro algoritmo, así como para mostrar los resultados obtenidos y poder compararlos con las propuestas de otros autores hemos elegido 12 bases de datos (ver Tabla 3.7) procedentes de datos reales. Excepto `Barbora`, el resto de los datos se pueden encontrar en el conocido repositorio de base de datos UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/).

La realización de los experimentos se han llevado a cabo en un Intel Core 2Duo de 1.73GHz con 1024MB de memoria principal, con Windows XP bajo el entorno

de desarrollo que ofrece NetBeans para Java. Además hemos implementado las propuestas de Suzuki y Hussain usando la formulación de la terna (csr, exc, ref) y $(csr, anom, ref)$ para la propuesta de Berzal et al. Todas ellas utilizando la computación mediante BitSets para comparar no los tiempos de extracción, si no el número de reglas anómalas y de excepción para ver la reducción obtenida en cada caso. También hemos impuesto que el número máximo de items en el antecedente o consecuente de la *csr* sea 2 en todas las bases de datos excepto en `Mushroom` donde lo hemos restringido a 1 debido al elevado número de *csr* encontradas.

Como anunciábamos al inicio de esta sección, el tiempo de ejecución está altamente influenciado por el número de reglas *csr* extraídas en el paso 2.1 del algoritmo y por el número de items puesto que hay que probar para cada *csr* y cada ítem si forman una regla de excepción. En la gráfica de la Figura 3.2, puede verse el tiempo de **ERSA** para las bases de datos `Mushroom, Abalone, Hepatitis, Barbora, Vote` y `Contraceptive` y los valores de $minsop \in \{0.05, 0.1, 0.2, 0.3, 0.4\}$ y $minconf \in \{0.95, 0.9, 0.75\}$ respectivamente. Para la base de datos `Nursery`, es necesario bajar el umbral de $minconf$ hasta 0.5 para obtener reglas common sense y por tanto para extraer las reglas de excepción asociadas. En la mencionada gráfica, puede verse claramente que aunque el número de transacciones entre `Mushroom` y `Barbora` es similar, debido al alto número de *csr* extraídas en el primer caso, el tiempo se ve incrementado. Aunque `Abalone` tiene menos transacciones que `Barbora`, **ERSA** emplea más tiempo en el primer caso porque tanto el número de *csr* como el de items es bastante mayor que en el segundo.

El tiempo de ejecución de **ERSA** ronda entre 1 segundo y 2 minutos en el resto de las bases de datos, haciendo que nuestro algoritmo pueda utilizarse en muy diversas bases de datos sin obtener tiempos superiores a la hora, siempre y cuando los umbrales *minsop* y $minconf$ estén ajustados para no sobrepasar un volumen muy alto de reglas *csr*.

Si comparamos los tiempos de ejecución de las distintas propuestas para extraer reglas de excepción usando **ERSA**, podemos observar que no hay grandes diferencias si utilizamos el soporte y la confianza. Recordemos que nosotros proponemos el uso del soporte y el factor de certeza sin utilizar la regla de referencia, y con ello bajamos considerablemente el tiempo de ejecución como podemos observar en la Figura 3.3 donde $minCF$ toma los mismos valores que $minconf$. Esta reducción es debida principalmente a que el número de *csr* disminuye cuando utilizamos el factor de certeza siendo las reglas common sense muy fuertes y las reglas de excepción fidedignas.

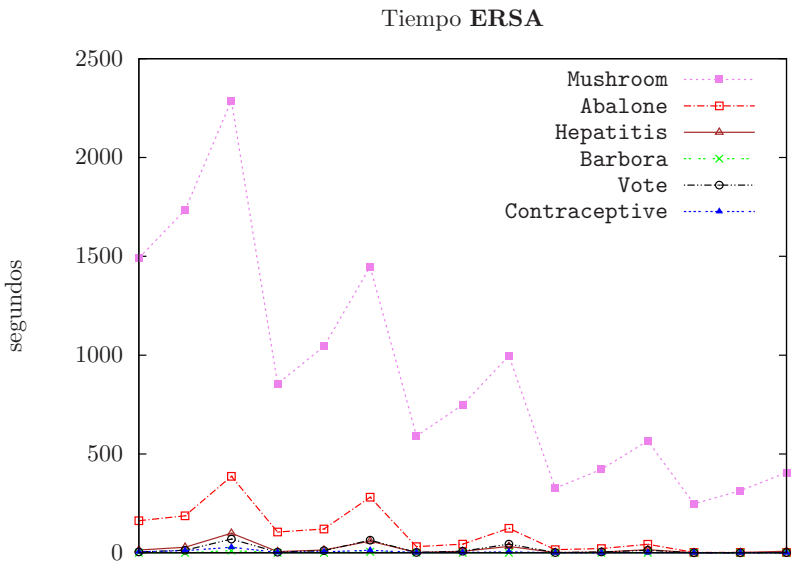

Figura 3.2: Tiempo en segundos para extraer excepciones del algoritmo *ERSA* usando diversos pares $(minsop, minconf)$.

En cuanto a la extracción de reglas anómalas, hemos comprobado que el comportamiento de **ERSA** y **ARSA** con respecto al tiempo es muy parecido variando en varios segundos de uno a otro según la base de datos y los umbrales elegidos. Como ejemplo ver la Figura 3.4 donde se muestran los tiempos de extracción de ambos algoritmos con la base de datos `Abalone`.

3.2.3. Comparación de Resultados

En las 12 bases de datos elegidas, podemos observar que, en términos generales, nuestro método obtiene resultados comparables a los obtenidos con la propuesta de Suzuki et al. aunque, en algunas de ellas el número de reglas de excepción se ve incrementado. Con respecto a la propuesta de Hussain et al. se encuentran un número muy bajo de reglas de excepción puesto que la imposición de la regla de referencia es una condición muy fuerte que nos decía que el ítem E causaba una excepción para dos reglas common sense. Observemos también en la Tabla 3.8 asociada a la base de datos `Mushroom`, que nuestra propuesta usando el factor de certeza disminuye el número de reglas de excepción con respecto al resto sin necesidad de imponer una regla de referencia. Sin embargo, en otras bases de datos,

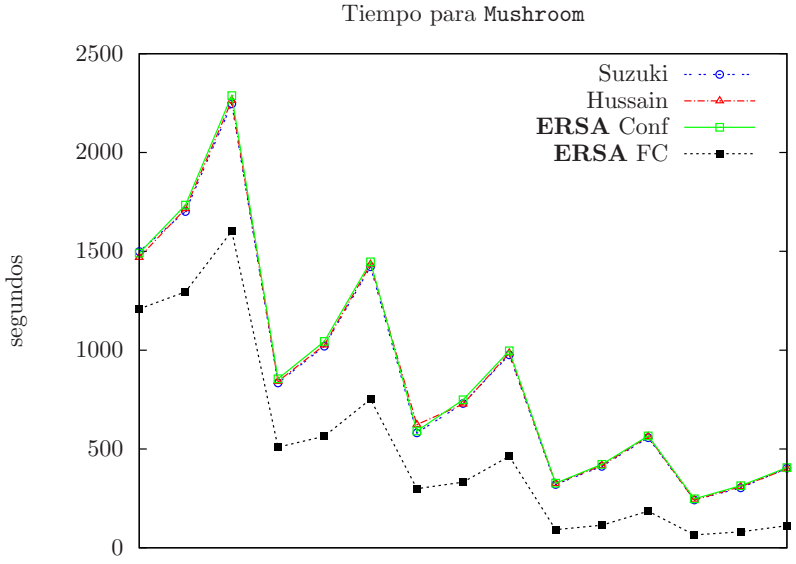

Figura 3.3: Tiempo en segundos para extraer excepciones con el algoritmo **ERSA** para las distintas propuestas en la base de datos `Mushroom`.

como en `Abalone` (Tabla 3.9) el número de reglas de excepción no es tan bajo. Cuando esto sucede, hemos observado que muchas de las *csr* tienen un elevado número de reglas de excepción asociadas. Esto podría evitarse imponiendo un criterio de poda para restringirnos sólo a aquellos pares (csr, exc) que no superen las 3 reglas de excepción asociadas. También sería interesante estudiar a qué se debe que dichas reglas de asociación tengan asociadas tantas reglas de excepción.

En las Tablas 3.10 y 3.11, podemos ver que el número de reglas de excepción disminuye en algunos casos para el enfoque de Suzuki et al. al disminuir el valor del umbral $minconf$. Esto es debido a que hemos fijado el mismo umbral tanto para la regla de excepción como para la regla de referencia. Cuando disminuimos el valor de $minconf$ la condición para la regla de referencia es más fuerte puesto que ésta exige que la regla $E \rightarrow \neg Y$ no sea confidente. En su enfoque, Suzuki et al. fijan distintos umbrales para cada una de las tres reglas de la terna (csr, exc, ref), pero nosotros, para simplificar, hemos considerado el mismo umbral para la confianza en las tres.

Para terminar la comparativa para el caso de las reglas de excepción entre nuestra propuesta y las otras, en las tablas anteriores se puede ver que nuestro

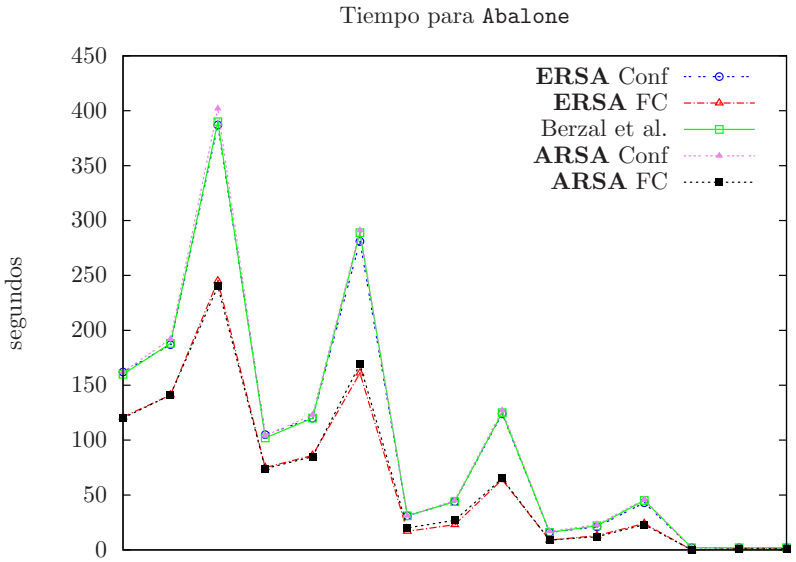

Figura 3.4: Tiempo en segundos para extraer excepciones y anomalías con el algoritmo **ERSA** y **ARSA** para las distintas propuestas en la base de datos `Abalone`.

Propuesta	$minconf/minFC$	$minsop = 0.05$		$minsop = 0.1$	
		$\#csr$	$\#exc$	$\#csr$	$\#exc$
Suzuki	0.95	384	212	221	117
	0.90	450	511	270	257
	0.75	594	1070	377	640
Hussain	0.95	384	10	221	0
	0.90	450	58	270	8
	0.75	594	312	377	98
ERSA Conf	0.95	384	223	221	118
	0.90	450	557	270	279
	0.75	594	1490	377	875
ERSA FC	0.95	307	478	132	100
	0.90	335	618	146	148
	0.75	415	1274	195	395

Tabla 3.8: Número de reglas *csr* y de excepción para las distintas propuestas en la base de datos `Mushroom`.

Propuesta	$minconf/minFC$	$minsop = 0.05$		$minsop = 0.1$	
		$\#csr$	$\#exc$	$\#csr$	$\#exc$
Suzuki	0.95	587	35	383	19
	0.90	694	43	449	19
	0.75	1436	84	1053	52
Hussain	0.95	587	0	383	0
	0.90	694	0	449	0
	0.75	1436	0	1053	0
ERSA Conf	0.95	587	62	383	29
	0.90	694	136	449	79
	0.75	1436	1170	1053	945
ERSA FC	0.95	438	44	279	24
	0.90	520	69	320	29
	0.75	896	450	597	286

Tabla 3.9: Número de reglas *csr* y de excepción para las distintas propuestas en la base de datos `Abalone`.

Propuesta	$minconf/minFC$	$minsop = 0.05$		$minsop = 0.1$	
		$\#csr$	$\#exc$	$\#csr$	$\#exc$
Suzuki	0.95	176	24	98	13
	0.90	381	57	203	27
	0.75	988	97	484	22
Hussain	0.95	176	0	98	0
	0.90	381	0	203	0
	0.75	988	11	484	1
ERSA Conf	0.95	176	24	98	13
	0.90	381	62	203	27
	0.75	988	225	484	88
ERSA FC	0.95	12	1	7	1
	0.90	28	5	17	4
	0.75	171	65	83	35

Tabla 3.10: Número de reglas *csr* y de excepción para las distintas propuestas en la base de datos `Contraceptive`.

Propuesta	$minconf/minFC$	$minsop = 0.05$		$minsop = 0.1$	
		$\#csr$	$\#exc$	$\#csr$	$\#exc$
Suzuki	0.95	394	316	335	280
	0.90	1047	540	959	505
	0.75	2167	233	1961	163
Hussain	0.95	394	0	335	0
	0.90	1047	0	959	0
	0.75	2167	1	1961	0
ERSA Conf	0.95	394	549	335	462
	0.90	1047	2035	959	1875
	0.75	2167	8640	1961	7971
ERSA FC	0.95	94	49	65	33
	0.90	271	321	208	224
	0.75	1092	2319	977	2114

Tabla 3.11: Número de reglas *csr* y de excepción para las distintas propuestas en la base de datos `Wisconsin breast cancer.`

método consigue en la mayoría de los casos una reducción considerable del número de reglas de excepción obtenidas sin necesidad de usar una regla de referencia (obsérvese la reducción de reglas de excepción entre usar confianza y factor de certeza), imponiendo una medida más fuerte y fiable que la confianza, habiendo elegido en nuestro caso el factor de certeza debido a sus buenas propiedades, ya analizadas en la Sección 2.1.6.

En las Figuras 3.5∼3.10 también podemos ver el comportamiento de nuestras propuestas usando el factor de certeza para obtener reglas anómalas y de excepción frente al número de reglas common sense obtenido. En la primera de ellas (`mushroom`) nuestra propuesta para las reglas de excepción es peor para los casos de $minFC = 0.75$ con $minsop = 0.1$ ó 0.05. Algo parecido ocurre para las bases de datos `abalone` y `wisconsin breast cancer`, donde nuestra propuesta es bastante peor que la de Suzuki et al. cuando $minFC = 0.75$. En estos casos, volvemos a encontrarnos que muchas de las *csr* llevan asociadas un gran número de excepciones. Con el criterio de poda que propusimos anteriormente podríamos obtener mejores resultados.

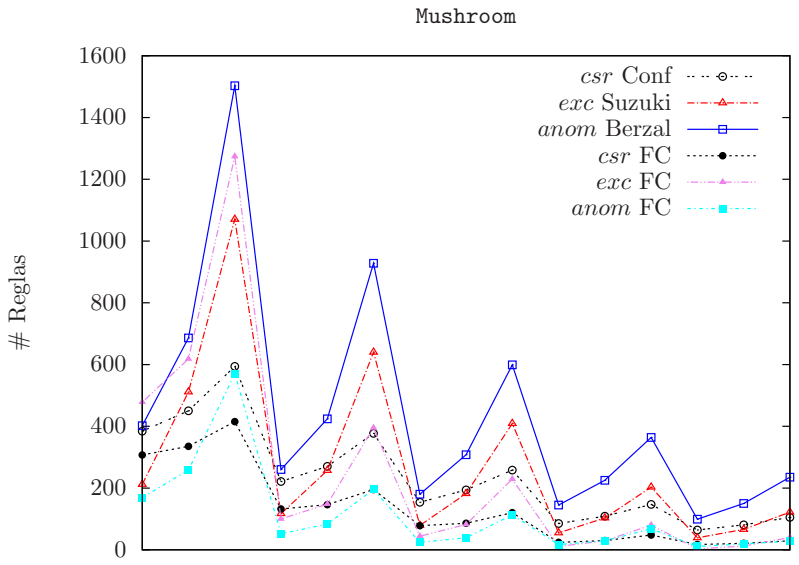

Figura 3.5: Número de reglas *csr*, de excepción y anómalas con las distintas propuestas para la base de datos `Mushroom`.

Con respecto a las reglas anómalas, en términos generales, siguiendo nuestra

propuesta extraemos un menor número de reglas que con el método de Berzal et al., obteniendo que el número de reglas anómalas extraídas es menor que el número de *csr* en la mayoría de las bases de datos.

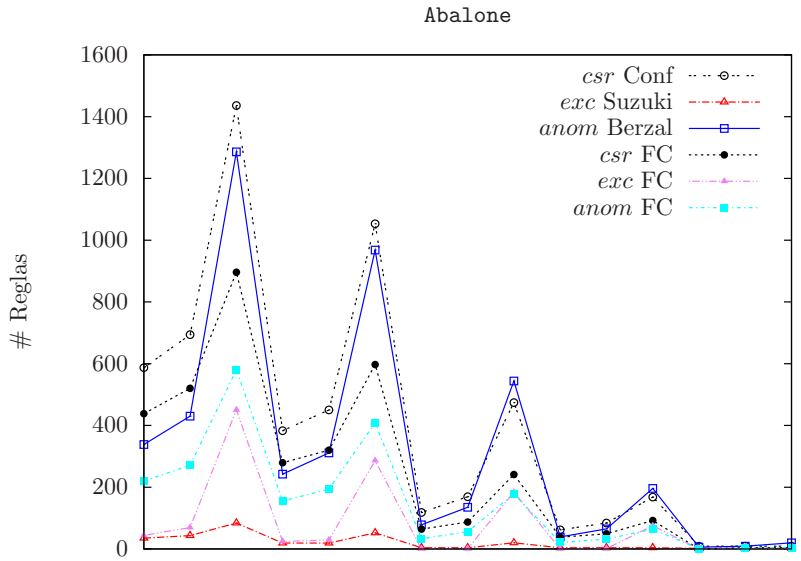

Figura 3.6: Número de reglas *csr*, de excepción y anómalas con las distintas propuestas para la base de datos `Abalone`.

En las Figuras 3.11 y 3.12 podemos ver el número de reglas dobles obtenidas en algunas de las bases de datos utilizando respectivamente la confianza y el factor de certeza como medida de interés para los umbrales de $minsop = 0.05$ y $minconf/minCF = 0.75$. En estos casos, la obtención de reglas dobles nos ofrece una información más detallada sobre la relación entre los items involucrados, puesto que éstos están estrechamente relacionados en ambos sentidos de la regla. Este tipo de regla brinda una nueva forma para ver si existe una mayor dependencia o una relación más fuerte entre un conjunto de items.

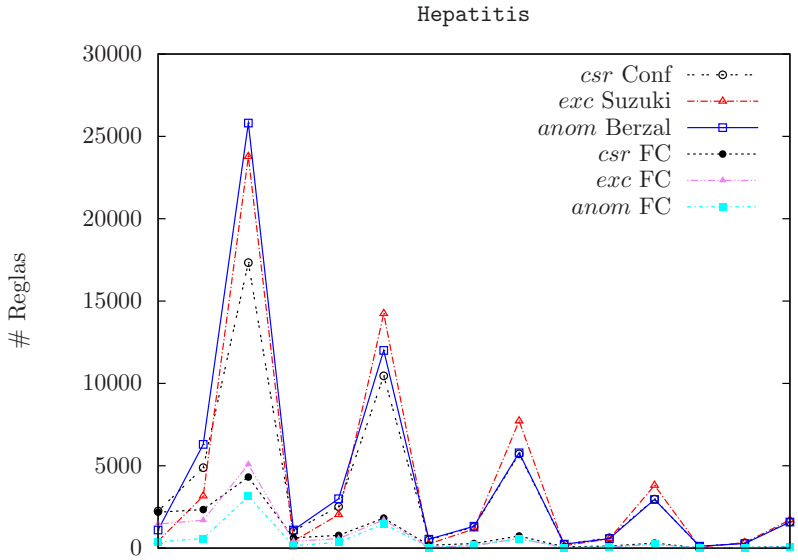

Figura 3.7: Número de reglas *csr*, de excepción y anómalas con las distintas propuestas para la base de datos `Hepatitis`.

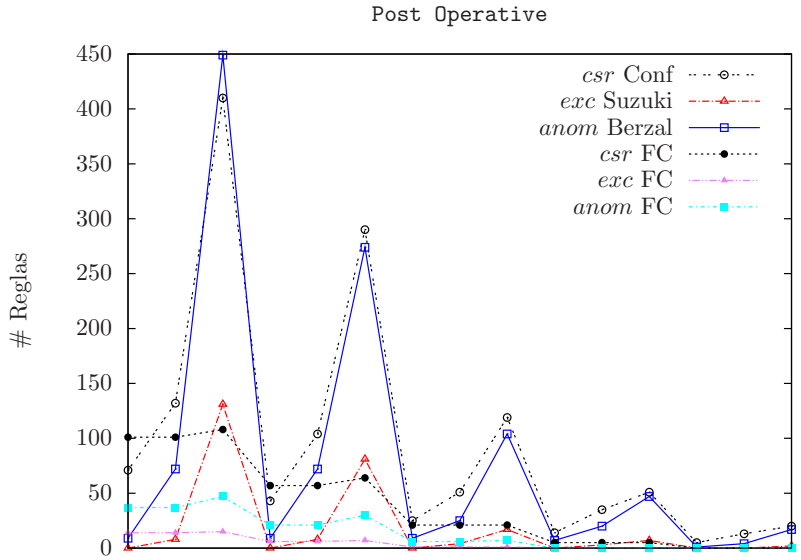

Figura 3.8: Número de reglas *csr*, de excepción y anómalas con las distintas propuestas para la base de datos `Post Operative`.

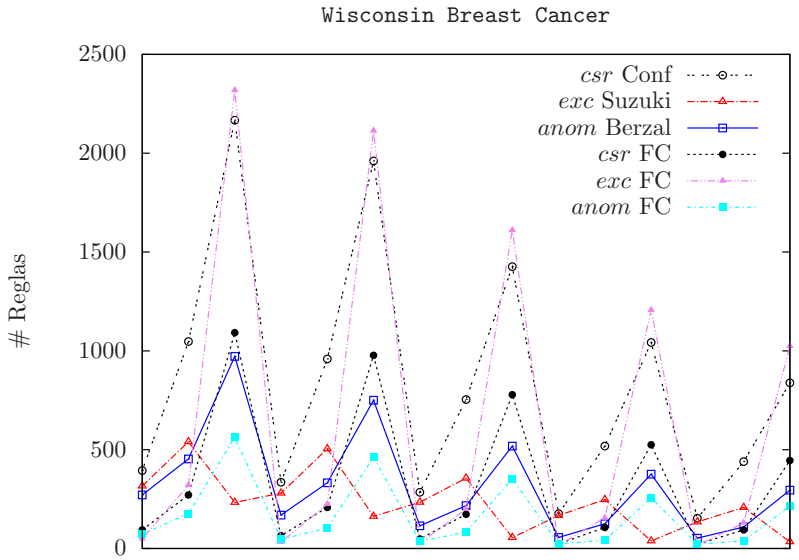

Figura 3.9: Número de reglas *csr*, de excepción y anómalas con las distintas propuestas para la base de datos `Wisconsin Breast Cancer`.

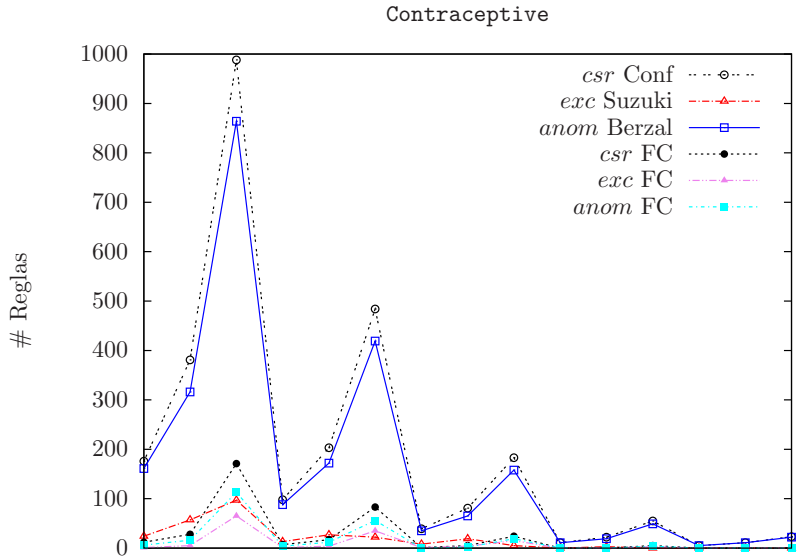

Figura 3.10: Número de reglas *csr*, de excepción y anómalas con las distintas propuestas para la base de datos `Contraceptive`.

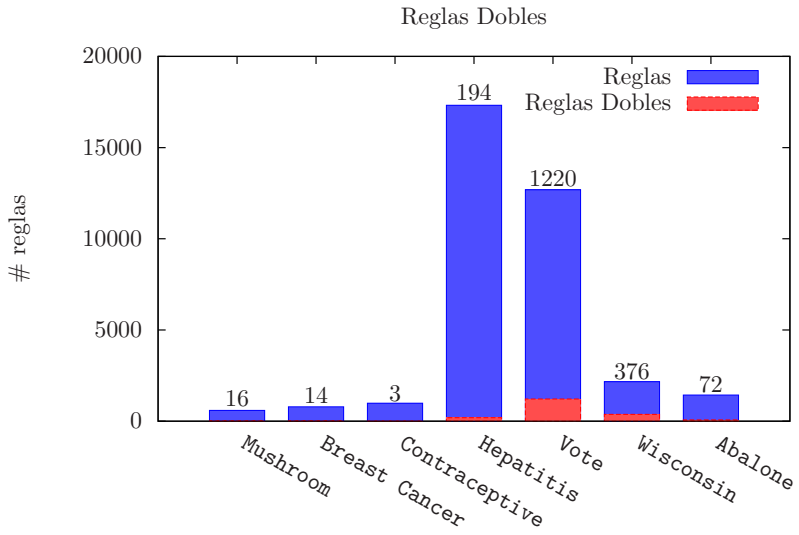

Figura 3.11: Número de reglas y reglas dobles para algunas de las bases de datos para $minsop = 0.05$ y $minconf = 0.75$.

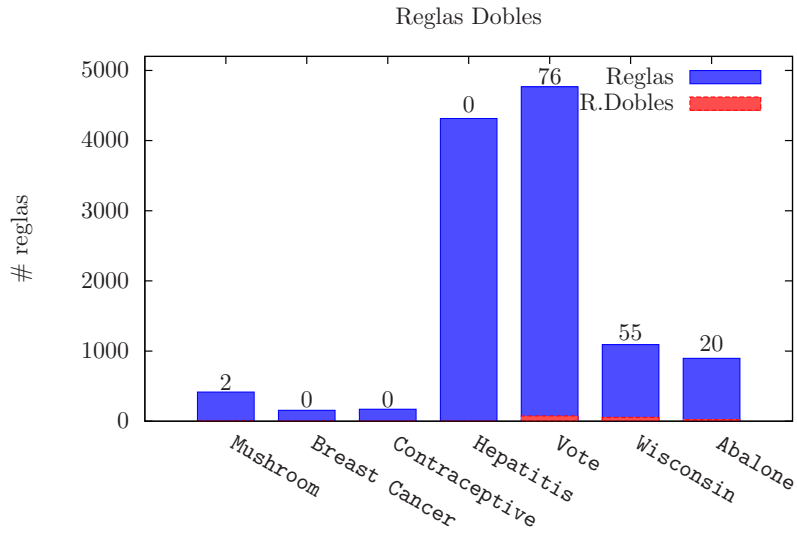

Figura 3.12: Número de reglas y reglas dobles para algunas de las bases de datos para $minsop = 0.05$ y $minFC = 0.75$.

3.2.4. Algunos Resultados Interesantes

Los experimentos realizados en la sección anterior con las distintas bases de datos nos ofrecen resultados que un experto en cada materia debe analizar para evaluar si realmente las reglas anómalas, de excepción y dobles obtenidas son útiles y novedosas.

Al no tener la opinión del experto, hemos tomado algunas de las reglas obtenidas y analizado si ofrecen información útil o interesante en algún aspecto. Las reglas que mostramos a continuación se han obtenido usando nuestra propuesta con el par (csr, exc) para las reglas de excepción y el triple (csr, $anom$, ref) para las reglas anómalas, imponiendo el factor de certeza.

En la base de datos `mushroom`, una de las reglas de excepción que nos podemos encontrar es la siguiente:

"SI `gill-color = brown`
$ENTONCES$ `edible` ($Sop = 0.114$ & $FC = 0.778$)
$EXCEPTO$ cuando `cap-color = pink` ($FC = 1$)".

Lo que viene a decir: que *si* el color del himenio (láminas debajo del sombrero) es marrón *entonces* el hongo es comestible con un soporte de 0.114 y un factor de certeza 0.778, *excepto* cuando el color del sombrero es rosa (con factor de certeza 1).

También podemos extraer algún par (csr,$anom$) de la base de datos `mushroom` como por ejemplo:

"SI `odor`=`foul`
$ENTONCES$ `ring-type = evanescent` ($Sop = 0.071$ & $FC = 1$)
$O\ BIEN$ `gill-size = broad` (de forma inusual con $FC_1 = 1$, $FC_2 = 1$),
$O\ BIEN$ `stalk-root = bulbous` (de forma inusual con $FC_1 = 1$, $FC_2 = 1$)".

La cual tiene dos reglas anómalas asociadas. Esta regla puede interpretarse como sigue: *si* el olor es parecido al del pescado *entonces* el tipo de anillo es evanescent con soporte 0.071 y factor de certeza 1, *o bien inusualmente* podemos obtener que el himenio tiene láminas anchas, *o bien*, la raíz del pie (o tallo) es bulboso o protuberante, con factor de certeza 1 tanto en la regla anómala como en la regla de referencia.

En los resultados obtenidos, hemos observado que muchas de las reglas anómalas obtenidas contienen itemsets complementarios al consecuente de la *csr*, es decir, A e Y tienen el atributo en común, pero difieren en el valor. Esto es muy útil para ver cuál es la asociación fuerte y cuáles sus comportamientos anómalos o consecuentes inusuales en este caso especial.

Para mostrar algún resultado sobre las reglas dobles, elegiremos dos obtenidas en la base de datos `vote` de los votos del Congreso de 1984 en los Estados Unidos:

$$\text{“ } \texttt{physician-fee-freeze = no} \longleftrightarrow \texttt{democrat}$$
$$(\ Sop = 0.563,\ FC_1 = 0.979\ \&\ FC_2 = 0.809)\text{”}$$

$$\text{“ } \texttt{el-salvador-aid = yes} \longleftrightarrow \texttt{democrat}$$
$$(\ Sop = 0.126,\ FC_1 = 0.979\ \&\ FC_2 = 0.809)\text{”}$$

que nos da una fuerte relación entre el voto a los demócratas y los dos antecedentes anteriores.

3.3. Resumen

Este capítulo ha mostrado la parte experimental relacionada con la implementación de un algoritmo viable y con muy buenas características para la extracción de reglas de excepción, anómalas y dobles. En particular se ha mostrado cómo hacer uso del modelo formal para obtener de la forma más eficiente y rápida la información sobre las frecuencias que involucran la medición de la validez de cada tipo de regla.

En particular, hemos utilizado un método basado en la representación de los items mediante `Bitsets` que acelera el cómputo de la operación lógica de la conjunción y la cardinalidad de los items y siguiendo un algoritmo un poco distinto a lo usual para obtener también las reglas dobles y las reglas de excepción y anómalas asociadas.

La complejidad de dicho algoritmo depende del número de reglas common sense extraídas y del número de items, como hemos podido observar en la Sección 3.2.2.

En general, los resultados obtenidos con nuestras propuestas son comparables a los obtenidos por Suzuki et al. para las reglas de excepción y mejoran bastante para las reglas anómalas en comparación con la propuesta de Berzal et al.

Con todo esto, también hemos visto que el modelo nos da una buen marco formal para la representación y la evaluación de distintos tipos de reglas de asociación y además ofrece una herramienta sencilla y útil a la hora de implementar los distintos tipos de cuantificadores para medir la validez de las reglas.

Conclusiones

A continuación resumiremos los objetivos conseguidos que se describieron al principio de la memoria y las conclusiones sobre los resultados obtenidos de nuestro trabajo.

El principal objetivo que propusimos al comienzo de esta memoria, era el **desarrollo de un modelo formal para la representación y evaluación de las reglas de asociación** que fuera extensible para cualquier tipo de asociación entre dos conjuntos de items u objetos, que sirviera de marco unificado para el estudio de las propiedades de la relación entre dos o más itemsets, así como de las medidas de interés utilizadas para su evaluación y extracción.

Este objetivo se ha alcanzado mediante el desarrollo de un modelo ya existente llamado GUHA [Hájek et al., 1966] que nos ha proporcionado una base lógica y estadística para avanzar y estudiar con más profundidad algunas de las herramientas que se utilizan en la extracción de reglas de asociación. Además hemos extendido el modelo a múltiples tipos de asociación como puede observarse en el Capítulo 2 en las Secciones 2.1, 2.2 y 2.4, donde hemos extendido el modelo para obtener reglas muy fuertes [Delgado et al., 2010b], reglas difusas [Delgado et al., 2009], reglas de excepción [Delgado et al., 2008a], [Delgado et al., 2010a], etc. definiendo en estos casos nuevos tipos de cuantificadores que se ajusten al tipo de conocimiento que se pretende extraer de la base de datos.

Al establecer dichas extensiones, se han obtenido herramientas y procedimientos que son de utilidad en diversas situaciones:

- Hemos propuesto un conjunto de axiomas o principios que toda medida de interés debería cumplir para obtener unos resultados deseables [Delgado et al., 2010b].

- Hemos establecido un vínculo entre los tipos de cuantificadores y las propiedades deseables para una "buena" medida de interés, proponiendo un procedimiento formal para comprobar cuando las medidas cumplen dichas propiedades por medio de su interpretación como cuantificador en el modelo formal (Secciones 2.1.4 y 2.1.7).

- Como resultado de lo anterior, hemos comprobado que la medida del factor de certeza presentado en [Berzal et al., 2002] cumple el conjunto de principios propuesto [Delgado et al., 2010b] utilizando como herramienta su expresión mediante el modelo y el procedimiento descrito en la Sección 2.1.4.

- En la Sección 2.2 hemos propuesto una extensión natural del modelo para el caso de reglas difusas utilizando el modelo de representación mediante niveles de restricción descrito en [Sánchez et al., 2009], comprobando que efectivamente cuando lo restringimos a reglas crisp, obtenemos el modelo que ya teníamos para este caso [Delgado et al., 2009].

- Como consecuencia de la extensión anterior, se ha obtenido un procedimiento para obtener medidas de interés para las reglas de asociación difusas a partir de cualquier medida de evaluación ya existente en el caso crisp.

- En la Sección 2.4 se han adaptado las dos herramientas básicas del modelo: tabla-4ft y cuantificadores-4ft para la extracción de tres tipos especiales de asociaciones: reglas de excepción, anómalas y dobles. Las dos primeras requieren la intervención de tres conjuntos de items involucrando la negación de algunos de ellos, obteniendo reglas que no corresponden al esquema habitual de *antecedente* $\rightarrow$ *consecuente*, si no al esquema *antecedente* $\wedge$ *excepción* $\rightarrow \neg$ *consecuente* para el caso de las reglas de excepción [Delgado et al., 2008a] o al esquema *antecedente* $\wedge \neg$ *consecuente* $\rightarrow$ *anomalía* para las reglas anómalas.

- La adaptación del modelo para estos tipos de reglas particulares ha ayudado a mejorar la implementación del algoritmo presentado en el Capítulo 3.

El segundo objetivo que nos propusimos fue **obtener nuevos procedimientos para obtener conocimiento de bases de datos formadas por bolsas**. Este objetivo queda totalmente desarrollado en la Sección 2.3, donde utilizando la Teoría de Subconjuntos Difusos y nuevos tipos de asociaciones tales como las dependencias graduales y las dependencias graduales difusas, hemos propuesto varias alternativas para descubrir conocimiento útil, comprensible y no

trivial de este tipo especial de datos que involucran cantidad. En particular se han desarrollado los siguientes objetivos:

- Se han examinado las propuestas existentes para este tipo de base de datos.
- Hemos presentado cuatro formas distintas de extraer diversos tipos de conocimiento. Hemos estudiado las características de cada propuesta y examinado la semántica obtenida con cada una, distinguiendo en qué casos es más útil una que otra.
- También hemos comparado nuestras propuestas con las ya existentes y hemos aplicado las propuestas a una base de datos ficticia.

Siguiendo en la línea de trabajo de obtener nuevos métodos para extraer conocimiento de utilidad y novedoso para el usuario se encuentra el tercer objetivo de nuestro trabajo donde se propuso **obtener y/o mejorar los procedimientos para la extracción de reglas de asociación que involucran la negación de items**. Este objetivo está ampliamente desarrollado en las Secciones 2.4 y 3.2. En la primera sección se aborda el problema desde un punto de vista formal y teórico haciendo uso del modelo, y en la segunda, de forma práctica mediante su implementación y experimentación con bases de datos reales. Concretamente podemos distinguir las siguientes aportaciones:

- Se ha realizado un estudio de los distintos tipos de reglas de asociación de este tipo, obteniendo que las reglas de excepción y las anómalas constituyen un buen punto de partida para obtener conocimiento de utilidad para el usuario.
- Se ha analizado la semántica y la formulación de las propuestas existentes para cada una de ellas.
- Hemos propuesto algunos cambios en la formulación de ambos tipos de reglas. Para las reglas de excepción, hemos visto que la regla de referencia puede restringir bastante la búsqueda. Para que esto no suceda hemos eliminado dicha restricción y hemos propuesto usar una medida de interés más fuerte y distinta a la confianza: el factor de certeza. Para las reglas anómalas, hemos cambiado la regla de referencia por una más estricta puesto que la anterior tenía problemas en este sentido, y hemos reforzado su extracción usando también el factor de certeza.

- Además, se ha presentado un nuevo tipo de reglas: las reglas dobles, que junto con la extracción de las reglas de excepción y las reglas anómalas asociadas, nos brinda una mejor descripción de la relación existente entre dos conjuntos de items.
- Se ha adaptado el modelo para la obtención de este tipo de reglas y para su posterior uso en la implementación práctica de las propuestas.

El último objetivo de esta memoria, como no podía ser de otra forma, ha sido el **desarrollo de un método, algoritmo o procedimiento basado en la filosofía del modelo para obtener distintos tipos de reglas de asociación**. En particular, el método presentado se ha desarrollado para obtener reglas de excepción, anómalas y dobles en una base de datos. Todo esto podemos verlo en el Capítulo 3 donde podemos destacar las siguientes contribuciones:

- Hemos analizado brevemente algunas de las propuestas para obtener reglas de asociación y hemos decidido utilizar una representación de los items mediante bitsets. Esta representación se ajusta muy bien al modelo cuando realizamos las operaciones lógicas entre distintos items: cardinal, conjunción, disyunción y negación. Aunque la negación es más costosa computacionalmente hablando, el modelo nos permite obtener las frecuencias necesarias por medio del resto de operaciones.
- Utilizando la anterior representación, hemos propuesto un algoritmo de extracción de reglas de asociación que puede utilizarse para cualquier tipo de cuantificador. Además lo hemos adaptado para extraer reglas dobles, de excepción y anómalas.
- Hemos estudiado la complejidad teórica y real con respecto al tiempo de ejecución y la memoria. Según este estudio, el tiempo de ejecución depende directamente del número de reglas common sense obtenidas y del número de items.
- Se han hecho experimentos con bases de datos reales y se han comparado los resultados con las otras propuestas para varias elecciones de los parámetros $minsop$ y $minconf/minCF$, obteniendo en muchos casos resultados muy prometedores.

En general, podemos decir que el trabajo realizado desarrolla varios aspectos fundamentales de la minería de datos, y en particular de la extracción de las reglas de asociación, algunos de ellos tenidos en muy pocas veces en consideración:

representación, formalización, desarrollo de nuevos procedimientos utilizando varios tipos de herramientas y experimentación en bases de datos reales con los métodos propuestos.

Trabajos Futuros

La principal dirección para futuros trabajos es investigar nuevos tipos de asociaciones que puedan ampliar el conocimiento del usuario. En esta línea ya hemos dado varios pasos con las aportaciones hechas en las Secciones 2.3 y 2.4 donde se ha obtenido un tipo de información más específica que el generalmente obtenido con simples reglas de asociación [Delgado et al., 2008c], [Delgado et al., 2008b], [Delgado et al., 2008a], [Delgado et al., 2010a]. Para ello hemos utilizado varias herramientas que pueden servir para futuras investigaciones: dependencias graduales y dependencias graduales difusas adaptadas a un tipo específico de base de datos, y la extracción conjunta de un par o una terna de reglas relacionadas entre sí cuyo conjunto ofrece una semántica de interés para el usuario.

Estas ideas pueden desarrollarse para otros casos y otros ambientes, para obtener información más compleja de las bases de datos. También es interesante el uso de la negación de un ítem para obtener nueva información o para confirmar la información ya obtenida, así como el papel que jugaba la regla de referencia en las ternas definidas para obtener reglas de excepción y anómalas que no fueran falsas u obtenidas por suerte.

Con respecto al modelo formal, este puede seguir extendiéndose y adaptándose a los nuevos tipos de asociaciones que surjan, proporcionando un marco formal único para el estudio de dichas asociaciones. En esta línea se encuentra el trabajo realizado para representar las reglas difusas [Delgado et al., 2009] descrito en la Sección 2.2 donde el modelo sirve como herramienta para definir nuevas medidas adaptadas a este tipo de reglas. A este respecto, esta línea de investigación puede seguir dando sus frutos.

Apéndice A

Teoría de Subconjuntos Difusos

La mayoría de la información que manejamos diariamente no es de tipo precisa o presenta imperfecciones. Entre los tipos de información imperfecta que nos podemos encontrar destacaremos los siguientes:

- **Incertidumbre.** Ocurre cuando no tenemos seguridad en la aparición de un cierto fenómeno o resultado de un cierto experimento. Por ejemplo: "mañana lloverá".
- **Imprecisión.** El valor de una variable se encuentra en un conjunto de valores pero no se puede precisar cuál es. Por ejemplo: "Juan tiene entre 20 y 25 años".
- **Vaguedad.** El conjunto que se especifica no está bien definido. Por ejemplo: "Juan es joven".

Además estos tipos de información pueden aparecer mezclados. Por ejemplo en la frase "creo que Juan gana mucho dinero" podemos encontrar imprecisión e incertidumbre ya que la información es imprecisa y no estamos seguros de que sea cierta.

Esto ha hecho que a lo largo de los años se hayan introducido modelos matemáticos para representar la información imperfecta[1], como la Teoría de la Probabilidad, la Teoría de la Evidencia de Dempster-Shafer y la Teoría de los Factores de Certeza. Para nuestros propósitos nos será de utilidad la

[1] Puede consultarse el capítulo 5 del libro [Lucas and van der Gaag, 1991] y el capítulo 7 del libro [Frost, 1987] para una breve descripción de algunas de estas teorías.

Teoría de Subconjuntos Difusos propuesta por Zadeh [Zadeh, 1965], la cual ha experimentado un considerable crecimiento debido a las aportaciones de otros muchos investigadores. En este documento, el tratamiento de la información es muy importante puesto que para la extracción de reglas de asociación podemos encontrarnos con varios tipos de datos que pueden contener algún tipo de imprecisión, vaguedad o incertidumbre. Pero ésta no será la única herramienta que utilizaremos para tales efectos puesto que el modelo de representación mediante *Niveles de Restricción* introducida por Sánchez et al [Sánchez et al., 2008b] y que desarrollamos en la sección 1.3 resuelve algunos de los problemas de la anterior cuando las operaciones involucradas contienen la negación, reduciendo además la complejidad cuando se ven envueltas un gran número de operaciones.

A.1. Concepto de Conjunto Difuso

El concepto clásico de conjunto está relacionado íntimamente con el cumplimiento de una determinada propiedad que deben satisfacer los elementos de ese conjunto. Podemos considerar que una propiedad es una función definida sobre un conjunto de objetos del referencial X, que hace corresponder cada uno de esos objetos con un elemento del conjunto $\{0, 1\}$. De tal forma, un objeto pertenecerá al conjunto si la función le asigna un 1; y en caso contrario, no pertenecerá a él. Estos conjuntos se denominan crisp o clásicos.

La Teoría de Subconjuntos Difusos generaliza esta idea clásica de conjunto, considerando que las propiedades que definen un conjunto son igualmente funciones definidas sobre el universo X, pero que usan como imagen el intervalo real cerrado $[0, 1]$. Con esta generalización se permite manejar la vaguedad o imprecisión que puede aparecer sobre los elementos del conjunto. De esta forma sabremos si un elemento pertenece o no a un conjunto difuso mediante un grado de aceptación entre 0 y 1.

Formalmente, sea X un universo (también llamado referencial). Se denomina *Subconjunto Difuso de* X a todo conjunto de la forma

$$A = \{(x, g), x \in X, g \in [0, 1])\}$$

es decir, A está formado por objetos del universo X y cada uno de ellos tiene asociado un grado de pertenencia medido en el intervalo $[0, 1]$.

El conjunto difuso A se caracteriza unívocamente por medio de la denominada *función de pertenencia*

$$\begin{aligned} \mu_A : X &\longrightarrow [0, 1] \\ x &\longmapsto \mu_A(x) = g \end{aligned}$$

de forma que para cada elemento x del conjunto referencial, μ_A representa el grado de pertenencia del elemento x al conjunto difuso A, tal y como puede verse en la figura A.1

Cuando el universo de referencia X es finito ($X = \{x_1, \ldots, x_n\}$) la función de pertenencia suele expresarse mediante la siguiente notación:

$$A = \{\mu_A(x_1)/x_1 + \mu_A(x_2)/x_2 + \cdots + \mu_A(x_n)/x_n\}$$

Definiremos a continuación algunos conceptos básicos sobre conjuntos difusos.

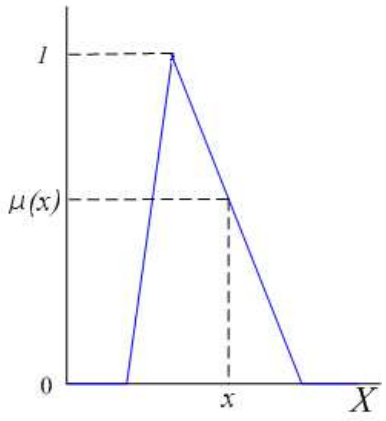

Figura A.1: Función de pertenencia de un conjunto difuso A

- Se dice que un conjunto difuso A es *normal* si existe al menos un $x \in X$ tal que $\mu_A(x) = 1$.
- Se llama *soporte* de un conjunto difuso A al conjunto de todos los elementos de X que tienen un nivel de pertenencia estrictamente mayor que 0, es decir:

$$\text{sop}(A) = \{x \in X \mid \mu_A(x) > 0\}.$$

- Se llama *núcleo* (kernel) del conjunto difuso A al conjunto

$$\text{Ker}(A) = \{x \in X \mid \mu_A(x) = 1\}$$

- El conjunto de todos los subconjuntos difusos sobre un referencial X se denota por $\tilde{\mathcal{P}}(X)$. Es obvio que

$$\mathcal{P}(X) \subset \tilde{\mathcal{P}}(X)$$

  donde $\mathcal{P}(X)$ denota las partes de X, es decir, el conjunto formado por todos los subconjuntos de X. Usaremos el término "crisp" cuando queramos diferenciar algún término "clásico" del caso difuso. En este caso $\mathcal{P}(X)$ es un concepto crisp mientras que $\tilde{\mathcal{P}}(X)$ es difuso.

- La *altura* de un conjunto difuso A se define como el mayor valor de su función de pertenencia:

$$Altura(A) = \sup_{x \in X} \mu_A(x)$$

A.2. Representación de Conjuntos Difusos mediante α-cortes

En la literatura pueden encontrarse diversas formas para representar un conjunto difuso sobre X de forma unívoca mediante un conjunto de subconjuntos crisp de X, cada uno de ellos con un cierto grado de importancia en la representación dado por un valor en el intervalo $[0, 1]$. La idea de representación es muy importante porque establece una conexión entre los conjuntos crisp y los difusos que permite extender propiedades de éstos primeros a los difusos. Aquí comentaremos la representación más utilizada de un conjunto difuso mediante el conjunto de sus conjuntos de nivel o de sus α-cortes.

Para cada α perteneciente al intervalo $[0, 1]$ y cada subconjunto difuso A, se define el α-corte (o corte de nivel α) de A como el conjunto de todos los elementos del universo de referencia que tienen un grado de pertenencia mayor o igual a α, es decir:

$$A_\alpha = \{x \in X \mid \mu_A(x) \geq \alpha\}.$$

Y el α-corte *estricto* (o fuerte) se define como

$$A^\alpha = \{x \in X \mid \mu_A(x) > \alpha\}.$$

Los α-cortes de un conjunto difuso guardan entre sí la relación de inclusión siguiente:

$$\text{Si } \alpha > \beta \text{ entonces } A_\alpha \subseteq A_\beta.$$

A partir de estas definiciones es inmediato comprobar que el soporte de un conjunto difuso A se corresponde con el 0-corte estricto de A y que el núcleo de un conjunto difuso A se corresponde con el corte de nivel 1.

El conjunto de valores $\alpha \in [0, 1]$ para los que existe al menos un elemento de X que pertenece al conjunto difuso A con grado α, se denomina *conjunto de niveles de* A, formalmente:

$$\Lambda(A) = \{\alpha \in [0, 1] \mid \mu_A(x) = \alpha \text{ para algún } x \in X\}.$$

El *Teorema de Descomposición* [Zadeh, 1971] establece que un conjunto difuso A puede representarse de forma unívoca a partir del conjunto de sus α-cortes. Formalmente:

$$A = \bigcup_{\alpha \in \Lambda(A)} \alpha A_\alpha,$$

donde $\bigcup$ representa la unión de conjuntos y αA_α es el conjunto difuso con función de pertenencia:

$$\mu_{\alpha A_\alpha}(x) = \begin{cases} \alpha & \text{para } x \in A_\alpha \\ 0 & \text{en otro caso.} \end{cases}$$

Es decir, $\mu_A(x) = \sup\{\alpha \mid x \in A_\alpha\}$.

Recíprocamente puede demostrarse que una familia finita de conjuntos $\{A_{\alpha_1}, \ldots, A_{\alpha_n}\}$ con pesos respectivos $\alpha_1, \ldots, \alpha_n$ que verifiquen que para cualesquiera $\alpha_i, \alpha_j \in [0, 1]$

$$\text{si } \alpha_i \geq \alpha_j \text{ entonces } A_{\alpha_i} \subseteq A_{\alpha_j},$$

define un conjunto de α-cortes del subconjunto difuso dado por

$$\mu_A(x) = \sup\{\alpha \mid x \in A_\alpha\}.$$

Ejemplo A.1. Sea $A = \{0.8/x_1,\ 0.4/x_2,\ 0.7/x_3,\ 0.6/x_4\}$. Entonces sus α-cortes asociados son:

$$\begin{aligned} A_{0.4} &= \{x_1, x_2, x_3, x_4\} \\ A_{0.6} &= \{x_1, x_3, x_4\} \\ A_{0.7} &= \{x_1, x_3\} \\ A_{0.8} &= \{x_1\} \end{aligned}$$

El anterior teorema nos dice que para cualquier $i = 1, 2, 3, 4$

$$\mu_A(x_i) = \sup\{\alpha \cdot \mu_{A_\alpha}(x_i)\},$$

siendo μ_{A_α} la función característica del α-corte A_α. ♦

Dado que una "relación" en sentido clásico no es más que un subconjunto de un producto cartesiano, es inmediato definir relaciones difusas.

Definición A.1. Una *relación difusa* R es un conjunto difuso definido sobre el producto cartesiano de universos de referencia: $\mu_R : X_1 \times \cdots \times X_n \longrightarrow [0, 1]$.

A.3. Operaciones básicas con conjuntos difusos

La extensión del concepto de conjunto para el caso difuso carecería de sentido si no extendiésemos también las operaciones que se pueden realizar entre conjuntos. Pero deben generalizarse de tal forma que cuando los conjuntos implicados sean los ordinarios, estas operaciones deben comportarse de la forma habitual.

Las familias de operadores difusos más importantes son las llamadas t-normas, t-conormas y negaciones, que nos servirán a su vez para extender los operadores de intersección, unión y complemento.

A.3.1. Operadores de intersección: t-normas

Una t-norma es una función $\otimes : [0,1]\times[0,1] \to [0,1]$ que verifica las propiedades siguientes para cualesquiera $a, b, c \in [0,1]$:

1. Frontera: $a \otimes 1 = a$.
2. Monotonía: Si $b \leq c$ entonces $a \otimes b \leq a \otimes c$.
3. Conmutatividad: $a \otimes b = b \otimes a$.
4. Asociatividad: $a \otimes (b \otimes c) = (a \otimes b) \otimes c$

Algunas de las t-normas más utilizadas son las siguientes:

- Intersección estándar: $a \otimes b = \text{mín}(a, b)$
- Producto algebraico: $a \otimes b = ab$
- Resta acotada (Pluralistic): $a \otimes b = \text{máx}(0, a + b - 1)$
- Intersección drástica: $a \otimes b = \begin{cases} a & \text{si } b = 1 \\ b & \text{si } a = 1 \\ 0 & \text{en otro caso.} \end{cases}$

La intersección de dos conjuntos difusos A y B mediante una t-norma se define como:

$$\mu_{A\cap B}(x) = \mu_A(x) \otimes \mu_B(x).$$

A.3.2. Operadores de unión: t-conormas

Una t-conorma es una función $\oplus : [0,1] \times [0,1] \longrightarrow [0,1]$ que verifica las cuatro propiedades siguientes:

1. Frontera: $a \oplus 0 = a$.
2. Monotonía: Si $b \leq c$ entonces $a \oplus b \leq a \oplus c$.
3. Conmutatividad: $a \oplus b = b \oplus a$.
4. Asociatividad: $a \oplus (b \oplus c) = (a \oplus b) \oplus c$.

Algunas de las t-conormas más utilizadas son las siguientes:

- Unión estándar: $a \oplus b = \text{máx}(a, b)$
- Suma algebraica: $a \oplus b = a + b - ab$
- Suma acotada (Pluralistic): $a \oplus b = \text{mín}(1, a + b)$
- Unión drástica: $a \oplus b = \begin{cases} a & \text{si } b = 0 \\ b & \text{si } a = 0 \\ 1 & \text{en otro caso.} \end{cases}$

La unión de dos conjuntos difusos A y B mediante una t-conorma se define como:

$$\mu_{A \cup B}(x) = \mu_A(x) \oplus \mu_B(x).$$

A.3.3. Operadores de complemento: negaciones

Una negación es una función $n : [0,1] \times [0,1] \longrightarrow [0,1]$ que verifica las dos propiedades siguientes:

1. Frontera: $n(0) = 1$ y $n(1) = 0$.
2. Monotonía: Si $a \leq b$ entonces $n(a) \geq n(b)$.

Estas propiedades son las mínimas que debe cumplir una negación, pero también se le suelen exigir algunas otras propiedades como:

3. Continuidad: n debe ser una función continua.

3. Propiedad involutiva: $n(n(a)) = a$ para todo $a \in [0,1]$.

La negación más utilizada es la negación estándar,

$$n(a) = 1 - a$$

aunque también hay otras como las llamadas negaciones umbrales,

$$n(a) = \begin{cases} 1 & a < umbral \\ 0 & a \geq umbral \end{cases}$$

donde $umbral \in]0, 1]$.

El complemento de un conjunto difuso A, que lo notaremos por $\bar{A}$ se definirá como

$$\mu_{\bar{A}}(x) = n(\mu_A(x)).$$

A.3.4. t-norma y t-conormas Duales

Dada una negación n, la t-norma $\otimes$ y la t-conorma $\oplus$ son duales con respecto a n si y sólo si

$$n(a \otimes b) = n(a) \oplus n(b) \text{ y } n(a \oplus b) = n(a) \otimes n(b)$$

A la terna $(\otimes, \oplus, n)$ se le denomina *triple dual*. Si consideramos la negación como $n(x) = 1 - x$, son triples duales las siguientes ternas:

- $(\text{mín}(a, b), \text{máx}(a, b), n)$
- $(ab, a + b - ab, n)$
- $(\text{máx}(0, a + b - 1), \text{mín}(1, a + b), n)$
- *(Intersección drástica, Unión drástica, n).*

También tienen especial relevancia y son especialmente interesantes las operaciones de inclusión e implicación de conjuntos difusos.

A.3.5. Inclusión e Implicación Difusas

Definición A.2. Sean A, B dos conjuntos difusos definidos sobre el mismo referencial X, diremos que A está incluido en B si su función de pertenencia toma valores más pequeños, esto es,

$$A \subseteq B \Leftrightarrow \mu_A(x) \leq \mu_B(x) \quad \forall x \in X$$

La anterior definición de L. Zadeh convierte la relación de inclusión en una pregunta crisp. Para evitar este inconveniente, se han propuesto numerosas definiciones alternativas para la relación de inclusión, como por ejemplo la siguiente, que utiliza una función de implicación $\Rightarrow$:

$$Inc(A, B) = \inf_{x \in X} A(x) \Rightarrow B(x)$$

Definición A.3. Una *implicación difusa* [Trillas and Valverde, 1985] es una función $i : [0,1] \times [0,1] \longrightarrow [0,1]$ que verifica las siguientes propiedades:

- Decreciente en primera variable: si $a \leq b$ entonces $i(a,x) \geq i(b,x)$ para cualesquiera $x, a, b \in [0,1]$.
- Creciente en segunda variable: si $a \leq b$ entonces $i(x,a) \leq i(x,b)$ para cualesquiera $x, a, b \in [0,1]$.
- Frontera: $i(0,x) = 1$, e $i(1,x) = x$ para todo $x \in [0,1]$.
- Asociativa: $i(a, i(b,x)) = i(i(b,a),x)$, para cualesquiera $x, a, b \in [0,1]$.

Entre los operadores de implicación destacan dos importantes subfamilias de operadores llamadas S-implicaciones y R-implicaciones.

Las S-implicaciones se definen mediante el uso de una t-conorma $\oplus$ y una negación n como

$$i(a,b) = n(a) \oplus b.$$

Algunas S-implicaciones son las siguientes:

- Kleene-Dienes: $i(a,b) = \text{máx}(1-a, b)$.
- Reichenbach: $i(a,b) = 1 - a + ab$.
- Pluralistic: $i(a,b) = \text{mín}(1, 1-a+b)$.
- Drástica: $i(a,b) = \begin{cases} b & a = 1 \\ 1-a & b = 0 \\ 1 & \text{en otro caso.} \end{cases}$

Las R-implicaciones se definen mediante el uso de una t-norma continua $\otimes$ según la expresión

$$i(a,b) = \sup\{x \in [0,1] : i(a,x) \leq b\}.$$

Algunas R-implicaciones son las siguientes:

- Gödel: $i(a,b) = \begin{cases} 1 & a \leq b \\ b & a > b. \end{cases}$.
- Goguen: $i(a,b) = \begin{cases} 1 & a \leq b \\ b/a & a > b. \end{cases}$.
- Pluralistic: $i(a,b) = \text{mín}(1, 1 - a + b)$.
- Drástica: $i(a,b) = \begin{cases} b & a = 1 \\ 1 & \text{en otro caso.} \end{cases}$

A.4. Números Difusos

Para poder estudiar la forma de operar con cantidades difusas necesitaríamos una Aritmética de las cantidades imprecisas, lo que se puede conseguir a partir de la siguiente definición de *número difuso*.

Un número difuso N es un conjunto difuso verificando las siguientes propiedades:

1. N es convexo, es decir, $\mu_N(\lambda x_1 + (1-\lambda)x_2) \geq \text{mín}\{\mu_N(x_1), \mu_N(x_2)\}$, para cualesquiera $x_1, x_2 \in [0,1]$,
2. N es normal,
3. μ_N es una función continua a trozos, y
4. el soporte de N está acotado.

Un número difuso A es de tipo LR si existen dos funciones de referencia L y R y dos números reales positivos α y β, tales que

$$\mu_A(x) = \begin{cases} L\left(\frac{a-x}{\alpha}\right) & x \leq a \\ R\left(\frac{x-a}{\beta}\right) & x \geq a \end{cases}$$

Al valor real a se le suele llamar moda de A, mientras que a α y β suelen denominarse amplitudes izquierda y derecha respectivamente.

Definición A.4. Un número difuso *triangular* T es un número difuso que se caracteriza mediante una terna de números reales (t_1, t_2, t_3) donde

- $t_1 \leq t_2 \leq t_3$
- $\mu_T(x) = \frac{x-t_1}{t_2-t_1}$ para cualquier $x \in [t_1, t_2]$.
- $\mu_T(x) = \frac{t_3-x}{t_3-t_2}$ para cualquier $x \in [t_2, t_3]$.
- $\mu_T(x) = 0$ para todo $x \in]-\infty, t_1[\cup]t_3, \infty[$

Definición A.5. Un número difuso *trapezoidal* Z es un número difuso que se caracteriza mediante una cuádrupla de números reales (z_1, z_2, z_3, z_4) donde

- $z_1 \leq z_2 \leq z_3 \leq z_4$
- $\mu_Z(x) = \frac{x-z_1}{z_2-z_1}$ para cualquier $x \in [z_1, z_2]$.

- $\mu_Z(x) = 1$ para cualquier $x \in [z_2, z_3]$.
- $\mu_Z(x) = \frac{z_4 - x}{z_4 - z_3}$ para cualquier $x \in [z_3, z_4]$.
- $\mu_Z(x) = 0$ para todo $x \in]-\infty, z_1] \cup [z_4, \infty[$

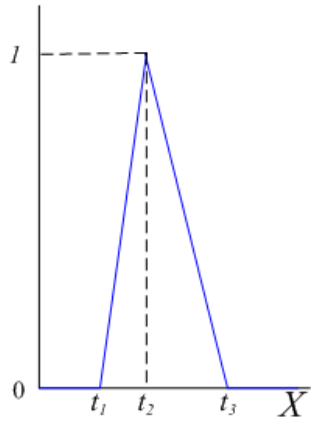

Figura A.2: Número triangular

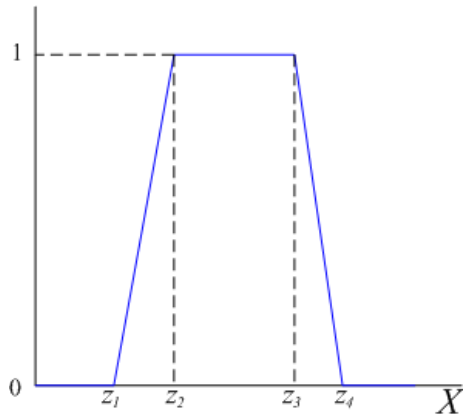

Figura A.3: Número trapezoidal

Como podemos observar, todo número difuso triangular (t_1, t_2, t_3) puede representarse como un número difuso trapezoidal (t_1, t_2, t_2, t_3).

A.5. Variables lingüísticas

En el ámbito de la lógica difusa, los conceptos imprecisos se representan mediante *etiquetas lingüísticas*. Estas son conjuntos difusos mediante los que se da semántica a un identificador que representa un concepto o una propiedad de este. Un ejemplo de etiqueta lingüística podría ser 'alto'y se representaría mediante un conjunto difuso que tiene como referencial al conjunto de todas las alturas posibles.

De manera informal, una *variable lingüística* es una variable que toma valores dentro de un dominio formado por etiquetas lingüísticas. Por ejemplo, la variable lingüística 'edad' de una persona podría tomar valores dentro del dominio {bebé, niño, adolescente, joven, maduro, mayor, muy mayor}. Estas etiquetas serán conjuntos difusos cuyo referencial es el conjunto (numérico) de las posibles edades de una persona.

El manejo de estas variables permite un razonamiento más cercano al del ser humano ya que disminuye la granularidad del dominio de una variable y además se definen significados intuitivos para la persona que los use.

Para cuantificar el grado de relación entre distintas variables lingüísticas en base a los datos almacenados en una base de datos, basta con medir los cardinales y los cardinales relativos de los conjuntos difusos inducidos por las etiquetas que compongan dichas variables y usarlos para la evaluación de sentencias cuantificadas. En la próxima sección veremos algunos métodos para la evaluación de dichas sentencias.

A.6. Cuantificación Difusa

El concepto de cuantificador lingüístico se debe a L. Zadeh [Zadeh, 1983]. El uso de cuantificadores lingüísticos en la lógica difusa es imprescindible para modelar sentencias del lenguaje natural llamadas sentencias cuantificadas, de las que trataremos posteriormente.

Dentro de los cuantificadores lingüísticos se suelen distinguir dos tipos:

- Absolutos, si se refieren a una cantidad o intervalos de cantidades; como por ejemplo: bastantes, pocos, muchos, alrededor de n, entre m y n, etc.
- Relativos, si representan porcentajes difusos del total; como por ejemplo: al menos la mitad, la mayoría, existe, aproximadamente la mitad, etc.

También es importante el concepto de *familia coherente de cuantificadores*, así una familia de cuantificadores relativos *coherentes* son aquellos que verifican las dos propiedades siguientes:

1. Condición de Frontera: $Q(0) = 0$ y $Q(1) = 1$;
2. Monotonía: si $x < y$ entonces $Q(x) \leq (y)$.

Según Zadeh [Zadeh, 1983] hay dos tipos de sentencias cuantificadas cuya estructura general es la siguiente:

- Sentencias de tipo I: Q de los X son A.
- Sentencias de tipo II: Q de los D son A.

donde Q es un cuantificador lingüístico, X es un conjunto finito crisp y tanto A como D son subconjuntos difusos de X que representan propiedades imprecisas. Las sentencias de tipo I son apropiadas tanto para cuantificadores absolutos como para relativos, mientras que las de tipo II sólo tienen sentido para los relativos.

La evaluación del grado de verdad de una sentencia cuantificada consiste en calcular la compatibilidad entre una medida de cardinalidad y un cuantificador.

La *cardinalidad* de un conjunto difuso trata de generalizar el concepto clásico de cardinal de un conjunto, esto es, el número de elementos que pertenecen al conjunto. Existen numerosas medidas para calcular la cardinalidad de un conjunto de difuso. Una de las más utilizadas es la medida de cardinalidad escalar.

Definición A.6. La *cardinalidad escalar* de un conjunto difuso A se define como:

$$|A| = \sum_{x \in X} \mu_A(x).$$

Lamentablemente, esta definición presenta varios problemas y conduce a numerosas paradojas. Una medida más acertada es la medida presentada en [Delgado et al., 2000], gracias a la cual la cardinalidad difusa es un conjunto difuso.

Definición A.7. La *cardinalidad difusa* de un conjunto difuso A es el conjunto difuso ED cuya función de pertenencia se evalúa para cada $0 \leq k \leq |\text{sop}(A)|$ como:

$$\mu_{ED}(A,k) = \begin{cases} \alpha_i - \alpha_{i-1} & \text{si } \alpha_i \in \Lambda(A) \text{ y } |A_{\alpha_i}| = k \\ 0 & \text{en otro caso} \end{cases}$$

donde $\Lambda(A) = \{\alpha_1, \ldots, \alpha_p\} \cup \{1\}$ son los α-cortes o niveles de A, siendo $\alpha_i > \alpha_{i-1}$ para todo $i = 1, \ldots, p$ y $\alpha_{p+1} = 1$.

Usando esta última definición, podemos determinar el cardinal relativo de un conjunto difuso con respecto a otro conjunto difuso.

Definición A.8. El *cardinal difuso relativo* de un conjunto difuso A con respecto a un conjunto B se define como un conjunto ER que se define para cada $0 \leq k \leq 1$ como:

$$ER(A/B,k) = \sum_{\alpha_i : C(A/B,\alpha_i) = k} \alpha_i - \alpha_{i-1}$$

donde

$$C(A/B,\alpha_i) = \frac{|A \cap B|}{|B|}$$

y donde $\Lambda(A) \cup \Lambda(A \cap B) = \{\alpha_1, \ldots, \alpha_p\}$, $\alpha_i > \alpha_{i+1}$ para cada $1 \leq i \leq p$, $\alpha_0 = 1$ y $\alpha_{p+1} = 0$.

Existen varios métodos para calcular el grado de cumplimiento de una sentencia cuantificada. Destacaremos el método de Zadeh para calcular el grado de cumplimiento de sentencias de tipo I y los métodos de Zadeh y el método GD para sentencias de tipo II.

Sentencias de tipo I

Recordemos que una sentencia de tipo I tiene la forma:

$$Q \text{ de los } X \text{ son } A.$$

Desglosaremos el funcionamiento general del método para el caso particular de la sentencia: "La mayoría de los alumnos son jóvenes", $Q = \textit{la mayoría}$, $X = \textit{alumnos}$ y $A = \textit{joven}$.

El método de Zadeh consta de los siguientes pasos:

1. Calculamos el cardinal de los alumnos jóvenes, es decir
$$|A| = \sum_{x \in X} \mu_A(x)$$
2. Calculamos el cardinal relativo de A con respecto a X, $|A| \,/\, |X| = \alpha_A$ donde $|X|$ es el cardinal de X, es decir, el número total de alumnos en este ejemplo.
3. Por último, el grado de cumplimiento de la sentencia según Zadeh sería:
$$Z_{mayoría} = Q_M(\alpha_A) = \alpha_A$$
donde $Q_M(x) = x$ es el cuantificador 'la mayoría'.

Sentencias de tipo II

Recordemos que una sentencia de tipo II tiene la forma:

$$Q \text{ de los } D \text{ son } A.$$

Consideramos la sentencia "La mayoría de los alumnos excelentes son jóvenes", $Q = \textit{la mayoría}$, $D = \textit{alumnos excelentes}$ y $A = \textit{alumnos jóvenes}$.

El método de Zadeh consta de los siguientes pasos:

1. Calculamos el cardinal de la intersección de los alumnos jóvenes con los alumnos excelentes, es decir
$$|A \cap D| = \sum_{x \in X} \mu_{A \cap D}(x) = \sum_{x \in X} \mu_A(x) \otimes \mu_D(x)$$
2. Calculamos el cardinal relativo de $A \cap D$ con respecto a X, $|A \cap D| \,/\, |X| = \alpha_{A \cap D}$ donde $|X|$ es el cardinal de X, es decir, el número total de alumnos en este ejemplo.
3. Por último, el grado de cumplimiento de la sentencia según Zadeh sería:
$$Z_{mayoría} = Q_M(\alpha_{A \cap D}) = \alpha_{A \cap D}$$
donde $Q_M(x) = x$ es el cuantificador 'la mayoría'.

El método GD [Delgado et al., 2000] permite calcular el grado de cumplimiento de una sentencia cuantificada de tipo II como:

- El grado de compatibilidad entre la medida de cardinalidad difusa relativa $ER(A/B)$ y Q, si Q es relativo.
- El grado de compatibilidad entre $ED(A/B)$ y Q, si Q es absoluto.

El método GD calcula el grado de cumplimiento de una sentencia de tipo II: "Q de los D son A" usando la siguiente fórmula

$$GD_Q\left(A/D\right) = \begin{cases} \sum\limits_{\alpha_i \in \Delta(A/D)} (\alpha_i - \alpha_{i+1}) Q\left(\dfrac{|(A \cap D)_{\alpha_i}|}{|D_{\alpha_i}|}\right) & \text{si } Q \text{ es relativo} \\ \sum\limits_{\alpha_i \in \Delta(A/D)} (\alpha_i - \alpha_{i+1}) Q\left(|(A \cap D)_{\alpha_i}|\right) & \text{si } Q \text{ es absoluto} \end{cases}$$

siendo $\Delta(A/D) = \Lambda(A \cap D) \cup \Lambda(D)$, y $\Delta(A/D) = \{\alpha_1, \ldots, \alpha_p\}$ con $\alpha_i > \alpha_{i+1}$ para todo $i \in \{1, \ldots, p\}$. Se presupone que D está normalizado, si no, se normaliza y el factor de normalización se aplica a $A \cap D$.

Bibliografía

[Aggarwal and Yu, 1998] Aggarwal, C. and Yu, P. (1998). A new framework for item set generation. In *Proceedings of the ACM PODS Symposium on Principles of Database Systems*, pages 18–24, Washington (USA).

[Agrawal et al., 1993] Agrawal, R., Imielinski, T., and Swami, A. (1993). Mining associations between sets of items in massive databases. In *ACM-SIGMOD International Conference on Data*, pages 207–216.

[Agrawal et al., 1996] Agrawal, R., Manilla, H., Sukent, R., Toivonen, A., and Verkamo, A. (1996). *Advances in Knowledge Discovery and Data Mining*, chapter Fast discovery of Association rules, pages 307–328. AAA Press.

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the Twentieth International Conference on Very Large Databases*, pages 487–499, Santiago, Chile.

[Au and Chan, 1997] Au, W. and Chan, K. (1997). Mining fuzzy association rules. In *Proc. 6th International Conference Information Knowledge Management*, pages 209–215.

[Bade et al., 2006] Bade, K., Hüllermeier, E., and Nürnberger, A. (2006). Hierarchical classification by expected utility maximization. In *Proceedings of the Sixth International Conference on Data Mining (ICDM'06)*.

[Balderas et al., 2005] Balderas, M., Berzal, F., Cubero, J., Eisman, E., and Marín, N. (2005). Discovering hidden association rules. In *KDD Workshop on Data Mining Methods for Anomaly Detection*, pages 13–20, Chicago.

[Bastide et al., 2000] Bastide, Y., Pasquier, N., Taouil, R., Stumme, G., and Lakha, L. (2000). Mining minimal non-redundant association rules using frequent closed itemsets. *CL 2000, LNAI 1861*, pages 972–986.

[Berzal et al., 2001] Berzal, F., Blanco, I., Sánchez, D., and Vila, M. (2001). A new framework to assess association rules. *LNCS 2189*, pages 95–104.

[Berzal et al., 2004] Berzal, F., Cubero, J., Marín, N., and Gámez, M. (2004). Anomalous association rules. In *IEEE International Conference on Data Mining*.

[Berzal et al., 2007] Berzal, F., Cubero, J., Sánchez, D., and Vila, M. (2007). An alternative approach to discover gradual dependencies. *Int. Journal of Uncertainty, Fuzziness and Knowledge-based Systems*, 15(5):559–570.

[Berzal et al., 2002] Berzal, F., Delgado, M., Sánchez, D., and Vila, M. (2002). Measuring accuracy and interest of association rules: A new framework. *Intelligent Data Analysis*, 6(3):221–235.

[Brijs et al., 2003] Brijs, T., Vanhoof, K., and Wets, G. (2003). Defining interestingness for association rules. *Information Theories & Applications*, 10:370–375.

[Brin et al., 1997] Brin, S., Motwani, R., Ullman, J., and Tsur, S. (1997). Dynamic itemset counting and implication rules for market basket data. *SIGMOD Record*, 26(2):255–264.

[Cai et al., 1998] Cai, C., Fu, A., Cheng, C., and Kwong, W. (1998). Mining association rules with weighted items. In *Proceedings of International Database Engineering and Applications Symposium (IDEAS 98)*.

[Chan et al., 2003] Chan, R., Yang, Q., and Shen, Y. (2003). Mining high utility itemsets. In *Proceedings of the third IEEE International Conference on Data Mining*.

[Chen and Wei, 2002] Chen, G. and Wei, Q. (2002). Fuzzy association rules and the extended mining algorithms. *Information Sciences*, 147:201–228.

[Chen and Huang, 2006] Chen, Y.-L. and Huang, T. C.-K. (2006). A new approach for discovering fuzzy quantitative sequential patterns in sequence databases. *Fuzzy Sets and Systems*, 157:1641–1661.

[Cock et al., 2005] Cock, M. D., Cornelis, C., and Kerre, E. (2005). Elicitation of fuzzy association rules from positive and negative examples. *Fuzzy Sets and Systems*, 149(1):73–85.

[Colton et al., 2005] Colton, S., Bundy, A., and Walsh, T. (2005). On the notion of interestingness in automated mathematical discovery. *International Journal of Human-Computer Studies*, 53:351–375.

[Davey and Priestley, 1994] Davey, B. A. and Priestley, H. A. (1994). *Introduction to Lattices and Order*. Cambridge University Press, fourth edition edition.

[Delgado et al., 2003a] Delgado, M., Marín, N., Sánchez, D., and Vila, M. (2003a). Fuzzy association rules: General model and applications. *IEEE Transactions on Fuzzy Systems*, 11(2):214–225.

[Delgado et al., 2003b] Delgado, M., Marín, N., Sánchez, D., and Vila, M. (2003b). Mining fuzzy association rules: An overview. *BISC International Workshop on Soft computing for Internet and Bioinformatics*, pages 351–374.

[Delgado et al., 2002a] Delgado, M., Martín-Bautista, M., Sánchez, D., Serrano, J., and Vila, M. (2002a). Association rule extraction for text mining. In *Proceedings of the 5th International Conference on Flexible Query Answering Systems*, pages 154–162.

[Delgado et al., 2002b] Delgado, M., Martín-Bautista, M., Sánchez, D., and Vila, M. (2002b). Mining text data: special features and patterns. In *ESF Exploratory Workshop on Pattern Detection and Discovery*, pages 140–153.

[Delgado et al., 2003c] Delgado, M., Martín-Bautista, M., Sánchez, D., and Vila, M. (2003c). On a characterization of fuzzy bags. *IFSA 2003, LNAI*, 2715:119–126.

[Delgado et al., 2008a] Delgado, M., Ruiz, M., and Sánchez, D. (2008a). Analyzing exception rules. In *IPMU'08*, pages 433–440, Málaga, Spain.

[Delgado et al., 2008b] Delgado, M., Ruiz, M., and Sánchez, D. (2008b). A logic approach for exceptions and anomalies in association rules. *Mathware & Soft Computing*, 15(3):285–295.

[Delgado et al., 2008c] Delgado, M., Ruiz, M., and Sánchez, D. (2008c). Pattern extraction from bag databases. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 16(4):475–494.

[Delgado et al., 2008d] Delgado, M., Ruiz, M., and Sánchez, D. (2008d). Reglas de asociación difusas: Nuevos retos. In *ESTYLF'08*, pages 523–528.

[Delgado et al., 2009] Delgado, M., Ruiz, M., and Sánchez, D. (2009). A restriction level approach for the representation and evaluation of fuzzy association rules. In *proceedings of the IFSA-EUSFLAT*, pages 1583–1588, Lisbon, Portugal.

[Delgado et al., 2010a] Delgado, M., Ruiz, M., and Sánchez, D. (2010a). *Analyzing Exception Rules*, volume 249 of *Studies in Fuzziness and Soft Computing*, pages 43–63.

[Delgado et al., 2010b] Delgado, M., Ruiz, M., and Sánchez, D. (Accepted, 2010b). Studying interest measures for association rules through a logical model. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*.

[Delgado et al., 2008e] Delgado, M., Ruiz, M., and Vila, M. (2008e). Un modelo lógico para la representación de reglas de asociación difusas mediante niveles de restricción. In *ESTYLF'08*, pages 371–376, Mieres, Spain.

[Delgado et al., 2000] Delgado, M., Sánchez, D., and Vila, M. (2000). Fuzzy cardinality based evaluation of quantified sentences. *Int. Journal of Approximate Reasoning*, 23:23–66.

[Ding and Yau, 2009] Ding, J. and Yau, S. S. (2009). TCOM, an innovative data structure for mining association rules among infrequent items. *Computers & Mathematics with Applications*, 57(2):290–301.

[Dong et al., 2007] Dong, X., Zheng, Z., and Niu, Z. (2007). Mining infrequent itemsets based on multiple level minimum supports. In *ICICIC '07: Proceedings of the Second International Conference on Innovative Computing, Informatio and Control*, pages 528–531.

[Dubois et al., 2003] Dubois, D., Hüllermeier, E., and Prade, H. (2003). A note on quality measures for fuzzy association rules.

[Dubois et al., 2006] Dubois, D., Hüllermeier, E., and Prade, H. (2006). A systematic approach to the assessment of fuzzy association rules. *Data Mining Knowledge Discovery*, 13(2):167–192.

[Dunham et al., 2000] Dunham, M., Xiao, Y., Gruenwald, L., and Hossain, Z. (2000). A survey of association rules. *http://www2.cs.uh.edu/ ceick/6340/grue-assoc.pdf*.

[Duval et al., 2007] Duval, B., Salleb, A., and Vrain, C. (2007). On the discovery of exception rules: A survey. *Studies in Computational Intelligence*, 43:77–98.

[Edmundson, 1969] Edmundson, H. (1969). New methods in automatic extracting. *Journal of the Association for Computing Machinery*, 16 (2):264–285.

[Fabris and Freitas, 1999] Fabris, C. and Freitas, A. (1999). Discovering surprising patterns by detecting occurrences of simpson's paradox. In *Development in Intelligent Systmes XVI*, pages 148–160, Cambridge, UK.

[Farzanyar et al., 2006] Farzanyar, Z., Kangavari, M., and Hashemi, S. (2006). A new algorithm for mining fuzzy association rules in the large databases based on ontology. *Proceedings of the Sixth IEEE Int. Conference on Data Mining-Workshops*, pages 65–69.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). *Advances in Knowledge Discovery and Data Mining*, chapter From Data Mining to Knowledge Discovery: An Overview, pages 1–34. AAAI/MIT Press.

[Fiot et al., 2008a] Fiot, C., Laurent, A., and Teisseire, M. (2008a). Fuzzy sequential pattern mining in incomplete databases. *Mathware and Soft Computing*, 15(1):41–59.

[Fiot et al., 2008b] Fiot, C., Masseglia, F., Laurent, A., and Teisseire, M. (2008b). Gradual trends in fuzzy sequential patterns. In *IPMU'08*, pages 456–463, Málaga, Spain.

[Frawley et al., 1992] Frawley, W., Piatetsky-Shapiro, G., and Matheus, C. (1992). Knowledge discovery in databases: An overview. *AAAI/MIT Press*, pages 57–70.

[Freitas, 1998] Freitas, A. (1998). A multi-criteria approach for the evaluation of rule interestingness. In *Proceedings of the International Conference on Data Mining*.

[Freitas, 1999] Freitas, A. (1999). On rule interestigness measures. *Knowledge-Based Systems*, 12:309–315.

[Frost, 1987] Frost, R. (1987). *Introduction to knowledge base systems*. Collins, London.

[Gago and Bento, 1998] Gago, P. and Bento, C. (1998). A metric for selection of the most promising rules. In *Proceedings of the 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD)*, pages 19–27.

[Gaines, 1991] Gaines, B. (1991). *Knowledge Discovery in Databases*, chapter The trade-off between knowledge and data in knowledge acquisition, pages 491–505. AAAI Press / MIT Press.

[Gaines, 1996] Gaines, B. (1996). Transforming rules and trees. *Advances in Knowledge Discovery and Data Mining. AAAI Press*, pages 205–226.

[Gamberger et al., 1999] Gamberger, D., Lavrac, N., and Jovanoski, V. (1999). High confidence association rules for medical diagnosis. In *Workshop of Intelligent Data Analysis in Medicine and Pharmacology.*

[Geng and Hamilton, 2006] Geng, L. and Hamilton, H. (2006). Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9.

[Goethals, 2003] Goethals, B. (2003). Survey on frequent pattern mining. *http://cchen1.csie.ntust.edu.tw/students/2009/Survey on frequent pattern mining.pdf.*

[Gyenesei, 2001] Gyenesei, A. (2001). A fuzzy approach for mining quantitative association rules. *Acta Cybern.*, 15(2).

[Hájek et al., 1966] Hájek, P., Havel, I., and Chytil, M. (1966). The GUHA method of automatic hypotheses determination. *Computing*, 1:293–308.

[Hájek and Havránek, 1978] Hájek, P. and Havránek, T. (1978). *Mechanising Hypothesis Formation-Mathematical Foundations for a General Theory.* Springer-Verlag: Berlin, Heidelberg and New York.

[Hájek et al., 2004] Hájek, P., Rauch, J., Coufal, D., and Feglar, T. (2004). The GUHA method, data preprocessing and mining. *LNAI*, 2682:135–153.

[Hand, 2002] Hand, D. (2002). Pattern detection and discovery. In *Pattern Detection and Discovery, LNAI 2447*, pages 1–12.

[Havrànek, 1974] Havrànek, T. (1974). Statistical quantifiers in observational calculi: An application in GUHA-methods. *Theory and Decision*, 6:213–230.

[Hébert and Crémilleux, 2006] Hébert, C. and Crémilleux, B. (2006). Optimized rule mining through a unified framework for interestingness measures. *LNCS*, 4081:238–247.

[Hilderman and Hamilton, 2001] Hilderman, R. and Hamilton, H. (2001). *Knowledge Discovery and Measures of Interest.* Kluwer Academic Publishers.

[Hilderman and Hamilton, 2003] Hilderman, R. and Hamilton, H. (2003). Measuring the interestingness of discovered knowledge: A principled approach. *Intelligent Data Analysis*, 7.

[Hipp et al., 2000] Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining: A general survey and comparison. *SIGKDD Explorations*, 2(1):58–64.

[Ho and Scott, 1996] Ho, K. and Scott, P. (1996). Zeta: a global method for discretization of continuous variables. In *Proceedings of KDD-96*, pages 44–49.

[Hong and Lee, 2008] Hong, T.-P. and Lee, Y.-C. (2008). *Fuzzy Sets and Their Extensions: Representation, Aggregation and Models*, chapter An Overview of Mining Fuzzy Association Rules, pages 397–410. Springer.

[Hong et al., 2006] Hong, T.-P., Lin, K.-Y., and Wang, S.-L. (2006). Mining fuzzy sequential patterns from quantitative transactions. *Soft Computing*, 10:925–932.

[Höppner, 2005] Höppner, F. (2005). Local pattern detection and clustering. *Local Pattern Detection, LNAI*, 3539:53–70.

[Houtsma and Swami, 1993] Houtsma, M. and Swami, A. (1993). Set-oriented mining for association rules in relational databases. *Technical Report RJ 9567.*

[Hsu et al., 2004] Hsu, P., Chen, Y., and Ling, C. (2004). Algorithms for mining association rules in bag databases. *Information Sciences*, 166:31–47.

[Hüllermeier, 2002] Hüllermeier, E. (2002). Association rules for expressing gradual dependencies. In *Proc. PKDD 2002 Lecture Notes in Computer Science, 2431*, pages 200–211.

[Hüllermeier, 2008] Hüllermeier, E. (2008). Fuzzy sets in machine learning and data mining. *Applied Soft Computing, In Press.*

[Hussain et al., 2000] Hussain, F., Liu, H., Suzuki, E., and Lu, H. (2000). Exception rule mining with a relative interestingness measure. In *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 86–97.

[Ikizler and Gvenir, 2001] Ikizler, N. and Gvenir, H. (2001). Mining interesting rules in bank loans data. In *Proceedings of the 10th Turkish Symposium on Artificial Intelligence an Neural Networks (TAINN)*, pages 238–246.

[Iványek, 1999] Iványek, J. (1999). On the correspondence between classes of implicational and equivalence quantifiers. In *PKDD'99, LNAI 1704*, pages 116–124.

[Jaroszewicz and Simovici, 2002] Jaroszewicz, S. and Simovici, D. (2002). Pruning redundant association rules using maximum entropy principle. *PAKDD'02, LNAI 2336*, pages 135–147.

[Jaroszewicz and Simovici, 2004] Jaroszewicz, S. and Simovici, D. (2004). Interestingness of frequent itemsets using bayesian networks as background knowledge. In *Proceedingsof the 2004, ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 178–186, Seattle, WA.

[Jia et al., 2003] Jia, L., Yao, J., and Pei, R. (2003). Mining association rules with frequent closed itemsets lattice. *KES 2003, LNAI 2773*, pages 469–475.

[Kamber and Shinghal, 1996] Kamber, M. and Shinghal, R. (1996). Evaluating the interestingness of characteristic rules. In *Proc.* 14^{th} *Int. Conf Knowledge Discovery and Data Mining*, pages 263–266.

[Karban, 2003] Karban, T. (2003). SDS-rules and classification. In *ECML/PKDD-2003 Discovery Challenge.*

[Karban et al., 2004] Karban, T., Rauch, J., and Šimunek, M. (2004). SDS-rules and association rules. In *ACM Symposium on Applied Computing*, pages 520–524, Nicosia, Cyprus.

[Kiran and Reddy, 2009] Kiran, R. U. and Reddy, P. K. (2009). An improved multiple minimum support based approach to mine rare association rules. In *IEEE Symposium on Computational Intelligence and Data Mining.*

[Kodratoff, 2001] Kodratoff, Y. (2001). Comparing machine learning and knowledge discovery in databases: An application to knowledge discovery in texts. In *Machine Learning and Its Applications*, pages 1–21.

[Korn et al., 2000] Korn, F., Labrinidis, A., Kotidis, Y., and Faloutsos, C. (2000). Quantifiable data mining using ratio rules. *The VLDB Journal*, 8:254–266.

[Kuok et al., 1998] Kuok, C., Fu, A., and Wong, M. (1998). Mining fuzzy association rules in databases. *SIGMOD Record*, 27(1):41–46.

[Lallich et al., 2006] Lallich, S., Teytaud, O., and E., P. (2006). Association rules interestingness: measure and validation. *Accepted in "Quality Measures in Data Mining".*

[Lee and Kwang, 1997] Lee, J. and Kwang, H. (1997). An extension of association rules using fuzzy sets. In *IFSA'97*, Prague, Czech Republic.

[Lin, 2000] Lin, T. (2000). Data mining and machine oriented modeling: A granular computing approach. *Applied Intelligence*, 13(2):113–124.

[Liu et al., 2000] Liu, B., Hsu, W., Chen, S., and Ma, Y. (2000). Analyzing the subjective interestingness of association rules. *IEEE Intelligent Systems*, 15(5):47–55.

[Liu et al., 1999a] Liu, B., Hsu, W., Mun, L., and Lee, H. (1999a). Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engeneering*, 11(6):817–832.

[Liu et al., 1999b] Liu, H., Lu, H., Feng, L., and Hussain, F. (1999b). Efficient search of reliable exceptions. In *Proceedings of the Third Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 194–203.

[López et al., 2007] López, F., Blanco, A., García, F., and Marín, A. (2007). Extracting biological knowledge by fuzzy association rule mining. *Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International*, pages 1–6.

[Louie and Lin, 2000a] Louie, E. and Lin, T. (2000a). A data mining approach using machine oriented modeling: Finding association rules using canonical names. In *Proceeding of 14th Annual International Symposium Aerospace/Defense Sensing, Simulation, and Controls*, volume 4057, Orlando.

[Louie and Lin, 2000b] Louie, E. and Lin, T. (2000b). Finding association rules using fast bit computation: Machine-oriented modeling. *LNAI*, 1932:486–494.

[Lucas and van der Gaag, 1991] Lucas, P. and van der Gaag, L. (1991). *Principles of expert systems*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Luhn, 1958] Luhn, H. (1958). The automatic creation of literature abstracts. *IBM Journal*, pages 159–165.

[Major and Mangano, 1995] Major, J. and Mangano, J. (1995). Selecting among rules induced from a hurricane database. *Journal of Intelligent Information Systems*, 4(1):39–52.

[Markou and Singh, 2003] Markou, M. and Singh, S. (2003). Novelty detection: a review-part 1: statistical approaches. *Signal Processing*, 83:2481–2497.

[Martinovic et al., 2008] Martinovic, J., Gajdos, P., and Snasel, V. (2008). Similarity in information retrieval. *CISIM*, 0:145–150.

[McGarry, 2005] McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *The Knowledge Engineering Review*, 20(1):39–61.

[Molina et al., 2008a] Molina, C., Sánchez, D., Serrano, J., and Vila, M. (2008a). Mining gradual dependencies with variation strength. *Mathware & soft computing*, 15:75–93.

[Molina et al., 2008b] Molina, C., Sánchez, D., Serrano, J., and Vila, M. (2008b). On some new types of fuzzy dependecies. In *IPMU'08*.

[Molina et al., 2009] Molina, C., Sánchez, D., Serrano, J., and Vila, M. (2009). Managing the absence of items in fuzzy association mining. In *proceedings of the IFSA-EUSFLAT*, pages 1571–1576, Lisbon, Portugal.

[Molina et al., 2007] Molina, C., Serrano, J., Sánchez, D., and Vila, M. (2007). Measuring variation strenght in gradual dependecies. In *EUSFLAT*, pages 337–344.

[Ohsaki et al., 2004] Ohsaki, M., Sato, Y., Kitaguchi, S., Yokoi, H., and Yamaguchi, T. (2004). Comparison between objective interestingness measures and real human interest in medical data mining. *IEA/AIE 2004, LNAI*, 3029:1072–1081.

[Ohshima et al., 2007] Ohshima, M., Zhong, N., Yao, Y., and Liu, C. (2007). Relational peculiarity-oriented mining. *Data Mining and Knowledge Discovery*, 15(2):249–273.

[Ordonez et al., 2000] Ordonez, C., Santana, C., and de Braal, L. (2000). Discovering interesting association rules in medical data. In *Proccedings of ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery*, pages 78–85.

[Padmanabhan, 2004] Padmanabhan, B. (2004). The interestingness paradox in pattern discovery. *Journal of Applied Statistics*, 31(8):1019–1035.

[Padmanabhan and Tuzhilin, 1998] Padmanabhan, B. and Tuzhilin, A. (1998). A belief driven method for discovering unexpected patterns. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 94–100.

[Padmanabhan and Tuzhilin, 1999] Padmanabhan, B. and Tuzhilin, A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27:303–318.

[Padmanabhan and Tuzhilin, 2002] Padmanabhan, B. and Tuzhilin, A. (2002). Knowledge refinement based on the discovery of unexpected patterns in data mining. *Decision Support Systems*, 33:309–321.

[Padmanabhan and Tuzhilin, 2006] Padmanabhan, B. and Tuzhilin, A. (2006). On characterization and discovery of minimal unexpected patterns in rule discovery. *IEEE Transactions on Knowledge and Data Engineering*, 18(2):202–216.

[Pasquier et al., 1998] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. (1998). Discovering frequent closed itemsets for association rules. *ICDT'99, LNCS 1540*, pages 398–416.

[Pei et al., 2000] Pei, J., Han, J., and Mao, R. (2000). CLOSET: An efficient algorithm for mining frequent closed itemsets. In *ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 21–30.

[Pei et al., 2004] Pei, J., Yin, Y., Mao, R., and Han, J. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery*, 8(1):53–87.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). Discovery, analysis and presentation of strong rules. *Knowledge Discovery in Databases*, pages 813–818.

[Piatetsky-Shapiro and Matheus, 1994] Piatetsky-Shapiro, G. and Matheus, C. (1994). The interestingness of deviations. In *Proceedings of AAAI Wordshop on Knowledge Discovery in Databases*.

[Rahal et al., 2008] Rahal, I., Rahhal, R., Wang, B., and Perrizo, W. (2008). CARIBIAM: Constrained association rules using interactive biological incremental mining. *Int. J. Bioinformatics Research and Applications*, 4(1):28–48.

[Ramaswamy et al., 1998] Ramaswamy, S., Mahajan, S., and Silberschatz, A. (1998). On the discovery of interesting patterns in association rules. In *Proceedings of the 24th International Conference on Very Large Data Bases*, pages 368–379, San Francisco: Kaufmann.

[Rauch, 2005] Rauch, J. (2005). Logic of association rules. *Applied Intelligence*, 22:9–28.

[Rauch and Šimunek, 2000] Rauch, J. and Šimunek, M. (2000). Mining for 4ft association rules. *Lecture Notes in Artificial Intelligence*, 1967:268–272.

[Rauch and Šimunek, 2005] Rauch, J. and Šimunek, M. (2005). An alternative approach to mining association rules. *Studies in Computational Intelligence (SCI)*, 6:211–231.

[Rocacher, 2003] Rocacher, D. (2003). On fuzzy bags and their application to flexible querying. *Fuzzy Sets and Systems*, 140:93–110.

[Roddick and Rice, 2001] Roddick, J. and Rice, S. (2001). What's interesting about cricket?-on thresholds and anticipation in discovered rules. *SIGKDD Explorations*, 3(1):1–5.

[Roddick and Spiliopoulou, 2002] Roddick, J. and Spiliopoulou, M. (2002). A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):750–767.

[Roubos et al., 2000] Roubos, J., Setnes, M., and Abonyi, J. (2000). *Learning fuzzy classification rules from data*. Springer-Verlag Berlin/Heidelberg.

[Roussinov and Zhao, 2003] Roussinov, D. and Zhao, J. (2003). Automatic discovery of similarity relationships through web mining. *Decision Support Systems*, 35:149–166.

[Ruiz and Bailón, 2008] Ruiz, M. and Bailón, A. (2008). Summarizing structured documents through a fractal technique. *Lecture Notes in Business Information Processing*, 12:328–340.

[Sahar, 2002] Sahar, S. (2002). On incorporating subjective interestingness into the mining process. In *Proceedings of the 2002, IEEE International Conference on Data Mining*, pages 681–684, Maebashi City, Japan.

[Sánchez, 1999] Sánchez, D. (1999). *Acquisition of Relationships Between Attributes in Relational Databases (in Spanish)*. Ph.d. thesis, Department of Computer Science and Artificial Intelligence, University of Granada, Spain.

[Sánchez et al., 2008a] Sánchez, D., Delgado, M., and Vila, M. (2008a). A restriction level approach to probability and statistics. In *ESTYLF'08*, pages 107–112, Asturias, Spain.

[Sánchez et al., 2008b] Sánchez, D., Delgado, M., and Vila, M. (2008b). A restriction level approach to the representation of imprecise properties. In *Int.*

Conference on Information Processing and Management of Uncertainty, pages 153–159, Málaga, Spain.

[Sánchez et al., 2008c] Sánchez, D., Delgado, M., and Vila, M. (2008c). RL-numbers: An alternative to fuzzy numbers for the representation of imprecise quantities. In *IEEE International Conference on Fuzzy Systems*, pages 2058–2065.

[Sánchez et al., 2009] Sánchez, D., Delgado, M., Vila, M., and Chamorro-Martínez, J. (2009). Representation of vagueness via restriction levels. the case of vague quantities. *Fuzzy Sets and Systems*, page Submitted.

[Sánchez et al., 2008d] Sánchez, D., Serrano, J., Blanco, I., and Martín-Bautista, M. (2008d). Using association rules to mine for strong approximate dependencies. *Data Mining & Knowledge Discovery*, 16(3):313–348.

[Sánchez et al., 2008e] Sánchez, D., Serrano, J., Vila, M., Delgado, M., Calero, G., Sánchez, J., and Aranda, V. (2008e). Building a fuzzy logic information network and a decision-support system for olive cultivation in andalusia. *Spanish Journal of Agricultural Research*, 6:252–263.

[Savasere et al., 1995] Savasere, A., Omiecinski, E., and Navathe, S. (1995). An efficient algorithm for mining association rules in large databases. In *Proceedings of the 21st Conference on Very Large Databases*, pages 432–444, Zürich, Switzerland.

[Shekar and Natarajan, 2004] Shekar, B. and Natarajan, R. (2004). A framework for evaluating knowledge-based interestingness of association rules. *Fuzzy Optimization and decision Making*, 3:157–185.

[Shen et al., 2002] Shen, Y., Yang, Q., and Zhang, Z. (2002). Objective-oriented utility-based association mining. In *Proceedings of the IEEE International Conference on Data Mining*, pages 426–433, Japan.

[Shortliffe and Buchanan, 1975] Shortliffe, E. and Buchanan, B. (1975). A model of inexact reasoning in medicine. *Mathematical Biosciences*, 23:351–379.

[Silberschatz and Tuzhilin, 1995] Silberschatz, A. and Tuzhilin, A. (1995). On subjective measures of interestingness in knowledge discovery. In *Proceedings of the 1st International Conferenceon Knowledge Discovery and Data Mining*, pages 275–281.

[Silberschatz and Tuzhilin, 1996] Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 8(6):970–974.

[Silverstein et al., 1998] Silverstein, C., Brin, S., and Motwani, R. (1998). Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2:39–68.

[Sim and Indrawan, 2007] Sim, A. T. H. and Indrawan, M. (2007). Importance of negative associations and mining of association pairs. In *Proceedings of iiWAS2007)*, pages 169–176.

[Sim et al., 2008a] Sim, A. T. H., Indrawan, M., and Srinivasan, B. (2008a). The importance of negative associations and the discovery of association rule pairs. *International Journal of Business Intelligence and Data Mining (IJBIDM)*, 3(2):158–176.

[Sim et al., 2008b] Sim, A. T. H., Indrawan, M., and Srinivasan, B. (2008b). Mining infrequent and interesting rules from transaction records. In *7th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Databases (AIKED'08)*, pages 515–520, UK.

[Singh et al., 2005] Singh, N., Singh, S., and Mahanta, A. (2005). CloseMiner: Discovering frequent closed itemsets using frequent closed tidsets. In *Proceedings of the fifth IEEE International Conference on Data Mining*.

[Sinoara and Rezende, 2006] Sinoara, R. A. and Rezende, S. O. (2006). A methodology for identifying interesting association rules by combining objective and subjective measures. *Inteligencia artificial: Revista Iberoamericana de Inteligencia Artificial*, 10(32):19–27.

[Smyth and Goodman, 1992] Smyth, P. and Goodman, R. (1992). An information theoretic approach to rule induction from databases. *IEEE Transactions on Knowledge and Data Engineering*, 4(4):301–316.

[Srikant and Agrawal, 1996] Srikant, R. and Agrawal, R. (1996). Mining quantitative association rules in large relational databases. In *ACM-SIGMOD Int. Conference Management Data*, pages 1–12.

[Sudkamp, 2005] Sudkamp, T. (2005). Examples, counterexamples, and measuring fuzzy associations. *Fuzzy Sets and Systems*, 149(1):57–71.

[Suzuki, 1996] Suzuki, E. (1996). Discovering unexpected exceptions: A stochastic approach. In *Proceedings of the fourth international workshop on RSFD*, pages 225–232.

[Suzuki, 1997] Suzuki, E. (1997). Autonomous discovery of reliable exception rules. In *Proceedings of KDD-97*, pages 259–262.

[Suzuki, 2001] Suzuki, E. (2001). In pursuit of interesting patterns wieh undirected discovery of exception rules. *Progress Discovery Science. Lecture Notes in Artificial Intelligence*, 2281:504–517.

[Suzuki, 2002] Suzuki, E. (2002). Undirected discovery of interesting exception rules. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(8):1065–1086.

[Suzuki, 2004] Suzuki, E. (2004). Discovering interesting exception rules with rule pair. In *Proc. Workshop on Advances in Inductive Rule Learning at PKDD-04*, pages 163–178.

[Suzuki, 2006] Suzuki, E. (2006). Data mining methods for discovering interesting exceptions from an unsupervised table. *J. UCS*, 12(6):627–653.

[Suzuki and Shimura, 1996] Suzuki, E. and Shimura, M. (1996). Exceptional knowledge discovery in databases based on information theory. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 275–278. AAAI Press.

[Suzuki and Zytkow, 2005] Suzuki, E. and Zytkow, J. (2005). Unified algorithm for undirected discovery of exception rules. *International Journal of Intelligent Systems*, 20:673–691.

[Tan and Kumar, 2000] Tan, P. and Kumar, V. (2000). Interestingness measures for association patterns: A perspective. In *Technical Report TR00-036, (KDD 2000 Workshop on Postprocessing in Machine Learning and Data Mining)*.

[Tan et al., 2002] Tan, P., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *SIGKDD'02*, Alberta, Canada.

[Tan et al., 2004] Tan, P., Kumar, V., and Srivastava, J. (2004). Selecting the right objective measure for association analysis. *Information Systems*, 29:293–313.

[Taniar et al., 2008] Taniar, D., Rahayu, W., Lee, V., and Daly, O. (2008). Exception rules in association rule mining. *Applied Mathematics and Computation*, 205:735–750.

[Trillas and Valverde, 1985] Trillas, E. and Valverde, L. (1985). *Management Decision Support Systems using Fuzzy Sets and Possibility Theory*, chapter On Implication and Indistinguishability in the Setting of Fuzzy Logic, pages 196–212. Verlag TÜW.

[Tsumoto and Tanaka, 1995] Tsumoto, S. and Tanaka, H. (1995). Automated discovery of functional components of proteins from amino-acid sequences based on rough sets and change of representation. In *Proceedings of KDD-95*, pages 318–324.

[Wang et al., 1998] Wang, K., Tay, S., and Liu, B. (1998). Interestingness-based interval merger for numeric association rules. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*, pages 2121–128.

[Yager, 1986] Yager, R. (1986). On the theory of bags. *Int. Journal of General Systems*, 13:23–37.

[Yao et al., 2005a] Yao, Y., Wang, F., and Wang, J. (2005a). 'rule + exception'strategies for knowledge management and discovery. *LNAI*, 3642:69–78.

[Yao et al., 2005b] Yao, Y., Wang, F., Zeng, D., and Wang, J. (2005b). Rule + exception strategies for security information analysis. *IEEE Intelligent Systems*, pages 52–57.

[Yao and Zhong, 1999] Yao, Y. and Zhong, N. (1999). An analysis of quantitative measures associated with rules. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 479–488.

[Yue et al., 2000] Yue, J. S., Tsang, E., Yeung, D., and Shi, D. (2000). Mining fuzzy association rules with weighted items.

[Zadeh, 1965] Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

[Zadeh, 1971] Zadeh, L. (1971). Similarity relations and fuzzy orderings. *Information Sciences*, 3(2):177–200.

[Zadeh, 1983] Zadeh, L. (1983). A computational approach to fuzzy quantifiers in natural languages. *Computing and Mathematics with Applications*, 9(1):149–184.

[Zaki, 2000] Zaki, M. (2000). Generating non-redundant association rules. In *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 34–43.

[Zaki, 2004] Zaki, M. (2004). Mining non-redundant association rules. *Data Mining and Knowledge Discovery*, 9:223–248.

[Zaki and Hsiao, 2005] Zaki, M. and Hsiao, C. (2005). Efficient algorithms for mining closed itemsets and their lattice structure. *IEEE Transactions on Knowledge and Data Engineering*, 17(4):462–478.

[Zaki et al., 1997] Zaki, M., Parthasarathy, S., Ogihara, M., and Li, W. (1997). New algorthms for fast discovery of association rules. In *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, pages 283–286, Newport Beach, California.

[Zembowicz and Zytkow, 1996] Zembowicz, R. and Zytkow, J. (1996). *Advances in Knowledge Discovery and Data Mining*, chapter From contingency tables to varios forms of knowledge in database, pages 39–81. AAAI Press / MIT Press, California.

[Zhong et al., 2001] Zhong, N., Ohshima, M., and Ohsuga, S. (2001). Peculiarity oriented mining and its application for knowledge discovery in amino-acid data. In *Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 260–269, London, UK.

[Zhong and Yao, 2003] Zhong, N. and Yao, Y. (2003). Peculiarity oriented multidatabase mining. *IEE Transactions on Knowledge and Data Engineering*, 15(4):952–960.

[Zhou and Yau, 2007a] Zhou, L. and Yau, S. (2007a). Association rule and quantitative association rule mining among infrequent items. In *MDM '07: Proceedings of the 8th international workshop on Multimedia data mining*, pages 1–9.

[Zhou and Yau, 2007b] Zhou, L. and Yau, S. (2007b). Efficient association rule mining among both frequent and infrequent items. *Computers & Mathematics with Applications*, 54(6):737–749.

Abstract and Conclusions

Nowadays the appearance of new devices for storing the information has experimented a high increase. At the same time several kinds of tools have been developed for obtaining knowledge from different sources available for the user. The information can come from very different sources and can also be structured in diverse ways. The huge diversity of storage formats and the large variety of their contents give us a source of wealth for extracting distinct types of knowledge useful for the user.

Databases are a very useful tool for storing data jointly with their associated relations. All the information that they collect can be used for developing several approaches for extracting useful, novel and interesting knowledge for users. These three characteristics are very important for discerning knowledge and reject the spurious one.

Knowledge Discovery deals with the development of tools and proceses from selection, processing, data cleaning, any *Data Mining* mechanism and it also includes the interpretation and evaluation steps. More concisely, one of the most used tools in *Data Mining* is that of association rules extraction due to their simplicity, manageable, and to their interpretability, but sometimes, it is misunderstood giving them a causal meaning.

The main topic of this work is to develop a formal model for a better comprehension of extraction and evaluation methods for association rules. Another important issue is to offer an unified framework for adapting the formal model to different approaches which use association rules. Moreover, some new proposals and techniques for obtaining new kinds of knowledge from diverse sources of information have been presented in the document.

Antecedents

The starting point of this work is a model called GUHA (General Unary Hypotheses Automaton) [Hájek et al., 1966] originated in the sixties for developing a method which allows computers generate and evaluate interesting hypothesis from a specific problem. GUHA uses the notion of hypothesis which is similar to that of association rule presented by Agrawal et al. in the nineties [Agrawal et al., 1993]. This resemblance allows to use some foundation concepts of GUHA for representing how the evaluation proceses of association rules works and it also offers a new point of view for their study.

Association rules measures the existing relation between a set of objects in a database by means of their frequency of appearance in the transactional records. Different approaches presented in this field give measures for the validity, quality and interestingness associated to the extracted rule, and all of them are in terms of such frequencies. There are few efforts on establishing a set of axioms or principles that must be fulfilled for these interestingness measures, and quite a few for giving a formal model for any related concept about association rules. Such a model is necessary for unifying the numerous types of associations and can be useful for developing new methods for extracting them. Nevertheless, there are several theories and models giving different ways to formulate new algorithms and new association rules, but they suffer from not collecting the predecessor approaches concerning association rules.

The two basic concepts for developing the GUHA model are called four fold table, $4ft$-table for short, and $4ft$-quantifier. $4ft$-table collects the four elementary frequencies involved in a general association rule $A \rightarrow B$ in a 2×2 matrix, and the $4ft$-quantifier is associated to the accomplishment measure. Depending on the properties fulfilled by the $4ft$-quantifier and on the way of applying it, we could manage with different types of associations interesting for the user.

Besides the model and its development, we will present some transversal approaches for obtaining new kinds of knowledge from a database. The first one concerns a method or process for obtaining associations in bag databases. Bags reference to a special type of transactions where the main difference lies on the quantities associated to objects. The associated quantities not only say if an item is in a transaction, they also count the number of appearances of it in the transaction. This extra information offered by bags can be beneficial for obtaining associations involving tendencies about increase or decrease of quantities associated to the item.

The second approach concerns the negation of itemsets for extracting association rules with a specific semantics useful for the user. This kind of rules are infrequent, so for finding them, there are some approaches using pairs or triples of rules. Depending on the negated item position in the rule, we can describe knowledge of two types: exceptional or anomalous with respect to an strong rule (frequent and confident rule). First works in this field were presented at the end of nineties and recently by some authors [Suzuki, 1996], [Hussain et al., 2000], [Berzal et al., 2004], and they will be the base of our approaches.

Objectives

The general aim of this dissertation is to develop a formal model for the representation and evaluation of association rules as a tool in a data mining process. For that, we have set some objectives and we have developed other objectives that are affine to the main purpose of any data mining process: to discover useful, novel and comprehensive knowledge for the user.

In accordance to this, the concrete objectives of this thesis are the following:

- Formal model development for the representation and evaluation of association rules.
  - To study existing models.
  - To set a model and adapt it for several types of association rules: crisp, fuzzy, exception and anomalous rules.
  - To study the model properties and to establish an unified model.
  - To link the model concepts with the type of association rule.
  - To obtain a set of properties, principles or axioms that a desirable interest measure for association rules must fulfils. To study this properties in terms of the formal model.
- To provide new methods or approaches for obtaining knowledge from bag databases.
  - To examine the existing approaches in this ambit.
  - To elaborate new extraction methods taking into account the associated quantity of objects in each transaction.
  - To analyze and study the advantages and drawbacks of the proposals.

  - To compare them with the existing approaches.

- To analyze the semantics and to find new methods for extracting association rules involving the negation of items.

  - To do a wide and rigorous study of the different approaches using the negation of items in association rules.
  - To see which approaches represent the semantics in a better way for the user.
  - To modify and/or propose new methods and strategies for improving the existing ones.
  - To use the model for their study and formalization.
  - To obtain a comparative study of the different approaches behavior.

- To study a method, heuristic or algorithm based on the formal model philosophy for obtaining different kinds of association rules.

  - To analyze the algorithms presented until now for extracting association rules.
  - To propose an algorithm for the extraction of association rules following the formal model philosophy.
  - To study the algorithm complexity and its performance in several real databases.
  - To modify the approach for some of the above rules: exception, anomalous and double rules.

Thesis Structure

This dissertation is structured as follows. First, after this introduction where the main objectives have been exposed, it is necessary to explain in more detail the ambient where we will work in the rest of the work as well as to define some preliminary concepts and to fixed some notation for the whole work.

First chapter called Preliminaries Concepts, shows a wide study of several models for representing association rules and a first approximation to the chosen one for developing it in the rest of the dissertation. This chapter also contains some tools such us fuzzy rules, dependence rules and gradual

dependencies (Sections 1.1.2 and 1.1.3) and the restriction level representation theory (Section 1.3) that we will use in some parts of the work.

Second chapter includes all theoretic results developed for covering the proposed objectives and this gives the chapter's name. We can distinguish four different sections. In Section 2.1 we study several criteria for measuring the rule's interestingness. Then, we propose and justify to add two new principles to the set of principles set by Piatetsky-Shapiro in [Piatetsky-Shapiro, 1991]. We also analyze the narrow relation between the principles and the type of $4ft$-quantifier in the formal model. We finish the section defining new quantifiers, in particular a new $4ft$-quantifier provided by the certainty factor measure presented in [Berzal et al., 2002], proving that it satisfies every principle proposed.

In Section 2.2 we follow extending the formal model for fuzzy rules. For that, we use the Restriction Level Representation Theory [Sánchez et al., 2009] presented in Chapter 1. Thanks to the generalization, we offer a method to extend crisp interest measures to the fuzzy case in a natural way.

In Section 2.3 we show several methods for extracting knowledge in bag databases. First we define the concepts of bag and fuzzy bag. We continue analyzing previous approaches and then we present ours. Our proposed methods use the Fuzzy Set Theory (see Appendix A) and two important tools: fuzzy rules and gradual dependencies [Sánchez et al., 2008d]. By means of gradual dependencies we achieve a very useful type of knowledge that offers a relation between the variation of object quantities stored in a bag database.

In Section 2.4 we study new methods of knowledge discovery using association rules. In particular, we center our attention to those which involve the items negation and contain a specific information useful for users: exception and anomalous rules. We deeply analyze previous approaches and we propose some modifications and new ways for extracting the same type of knowledge. We also formalize the ideas using the formal model and we compare our approaches with the existing ones. We finish proposing a new type of rule called *double* that jointly with exception and anomalous rules we provide a more complete description of existing relation between two sets of items.

Last chapter shows a brief resume about main rule extraction algorithms stopping in one approach which uses a binary representation of items [Louie and Lin, 2000b] that will be useful to develop a method for extracting association rules using the GUHA model. In particular, we use this approach for extracting exception, anomalous and double rules, comparing our results with

previous approaches seen in Section 2.4.

The dissertation finishes with the conclusions about the presented work and some proposals for future lines in our research.

Conclusions

In the following we resume the achieved objectives described at the beginning of the dissertation and the conclusions about the obtained results.

The main scope proposed was **to develop a formal model for the representation and evaluation of association rules** that was extendable to every type of association between two sets of items or objects and which gives a unified framework to study the properties associated to the rules and the evaluation measures involved in their extraction process.

This objective has been achieved developing an existing model called GUHA [Hájek et al., 1966] which provides a logic and statistic base to analyze more deeply some tools used in the association rule mining process. In addition, we have extended the model to several types of rules as can be seen in Chapter 2 in Sections 2.1, 2.2 and 2.4, where we expand the model to obtain very strong rules [Delgado et al., 2010b], fuzzy rules [Delgado et al., 2009], exception rules [Delgado et al., 2008a], [Delgado et al., 2010a], etcetera, defining in these cases new quantifiers that adjust to the type of extracted knowledge.

When doing these extensions, we have obtained new tools and processes which can be useful in several situations:

- We have proposed a set of desirable axioms or principles that every interest measure must fulfil [Delgado et al., 2010b].
- We have fixed a link between the type of quantifier in the model and these desirable set of principles for a "good" interest measure, proposing a formal method for proving when a measure fulfils those properties by means its interpretation as a quantifier in the formal model (Sections 2.1.4 and 2.1.7).
- As a result of the previous link, we proved that certainty factor [Berzal et al., 2002] fulfils the set of proposed principles [Delgado et al., 2010b] using as a tool its representation as a $4ft$-quantifier in the formal model in Section 2.1.4.

- In Section 2.2 we proposed a natural extension of the model for the case of fuzzy rules using the Restriction Level Representation Theory presented in [Sánchez et al., 2009], proving that if it is restricted to crisp rules we obtain the original model [Delgado et al., 2009].
- As a consequence, we have developed a method for extending every crisp interest measure to the fuzzy case preserving its properties.
- In Section 2.4 we present the model adaptation for extracting three special types of association rules: exception, anomalous and double rules. In the first two types, the approaches involved three distinct items using the items negation, obtaining rules that does not correspond to the usual scheme $antecedent \rightarrow consequent$. They correspond to one of these schemas: $antecedent \wedge exception \rightarrow \neg\ consequent$ in the case of exception rules [Delgado et al., 2008a] or to $antecedent \wedge \neg\ consequent \rightarrow anomaly$ for anomalous rules.
- The model extension to these kind of rules has helped to improve the algorithm implementation seen in Chapter 3.

Our second scope was **to obtain new methods for extracting knowledge from bag databases**. This objective is completely developed in Section 2.3 where using the Fuzzy Sets Theory and new kinds of associations such us crisp and fuzzy gradual dependencies, we proposed several approaches for obtaining useful, comprehensive and non trivial knowledge involving the associated quantities and their variation. More precisely we have achieved the following objectives:

- We examined the proposed methods for bag databases.
- We present four distinct ways for extracting different types of knowledge, studying their properties and associated semantics. We also distinguish different situations where each approach is more convenient than others.
- We compare our approaches with the existing ones, applying our approach on a toy bag database.

Our third scope is also in the previous research line: obtaining new methods for extracting useful and novel knowledge but in this case, **involving items negation**. This objective is widely developed in Sections 2.4 and 3.2. The former tackle the problem from a theoretic point of view using the model, and the latter by an algorithm implementation for extracting knowledge from real databases. In particular, we can specify the following contributions:

- We studied different approaches using the items negation, concluding that exception and anomalous rules are a good starting point for obtaining useful knowledge for the user.
- We analyzed semantics and formulation of existing approaches.
- We proposed some changes in previous approaches and we proposed some improvements. For exception rules, we showed that the reference rule can be very restrictive. We suggested to drop this rule and instead, to use a stronger interest measure: the certainty factor. For anomalous rules, the authors had the opposite problem: too many rules. For this, we suggested to change the reference rule for a stronger one and to use the certainty factor instead of confidence.
- In addition, we present a new kind of association called double rule which is useful for describing more completely the existing relation between two sets of items using them in conjunction with exception and anomalous rules.
- We extended the model for obtaining this type of rules and we use it for the posterior algorithmic implementation.

Last scope of this dissertation, as usual in data mining research, was **to develop an algorithm or process based on the model philosophy for obtaining association rules described in the formal model**. This topic is contained in Chapter 3 where we apply the presented algorithm for obtaining exception, anomalous and double rules. The process followed was the following:

- We first analyzed some approaches for obtaining association rules, choosing an item representation by sets of bits. This representation fits very well to the model when we do logic computations such us conjunction, disjunction and negation. Cardinality is also easy to compute. Negation is computationally more expensive than rest of computations but we can use the $4ft$-table for obtaining the necessary frequencies.
- Using the representation by sets of bits, bitsets for short, we propose an extraction algorithm adaptable for every kind of $4ft$-quantifier. We also do some experiments for the particular cases of mining exception, anomalous and double rules.
- We study the theoretic and real time complexity of proposed algorithm and their memory requirements, obtaining that for exception and anomalous rules

the execution time depends directly on the number of common sense rules and the number of items.

- We performed a battery of experiments with real databases comparing results with previous approaches and obtaining very promising results depending, of course, on parameters $minsop$ and $minconf/minCF$ election.

In general, the presented work develops some fundamental aspects of data mining, in particular, in association rule mining, some of them very few times under consideration: representation, formalization, development of new methods combining several tools and experimentation in real databases.

Future Works

The main direction for future works is following the research on new types of associations for expanding the user's knowledge. In this line we have done some small steps with the contributions made in Sections 2.3 and 2.4 where we obtained a specific piece of information using simple association rules [Delgado et al., 2008c], [Delgado et al., 2008b], [Delgado et al., 2008a], [Delgado et al., 2010a]. For that we used several tools that might be interesting for future works: crisp and fuzzy gradual dependencies adapted to a particular type of database, and the conjunct extraction of a set of rules, in our work, pairs and triples of rules defined for offering a semantic interpretation with interesting semantics for the user.

The same ideas can be suitable for obtaining more complex knowledge from databases. It is also interesting the use of the item negation for obtaining new information or for confirming the validity of obtained one as the role of the reference rule when mining exception or anomalous rules that dropped spurious rules.

With respect to the formal model, it could be adapted to new types of associations, offering an unified framework for studying their properties and for developing new extraction methods. In this line it is the work described in Section 2.2 for representing fuzzy rules where the model serves as a tool for defining new validity measures for fuzzy rules [Delgado et al., 2009]. This research line can be a fruitful one too.