Tratamiento de contextos con valores ausentes*

Cristina Alcalde
Dept. Matemática Aplicada
Universidad del País Vasco
Plaza de Europa,1
20018 San Sebastián
mapalvac@sp.ehu.es

Ana Burusco, Ramón Fuentes-González
Dept. Automática y Computación
Universidad Pública de Navarra
Campus de Arrosadía
31006 Pamplona
{burusco,rfuentes}@si.unavarra.es

Resumen

El objetivo de este trabajo es encontrar herramientas que nos permitan extraer información de un contexto L-Fuzzy intervalo-valorado en el cual existen elementos que son desconocidos.

Utilizando la idea de conjuntos frecuentes que aparece en minería de datos daremos la definición de conjunto L-fuzzy intervalo-valorado frecuente y utilizaremos esta definición para decidir en qué casos es posible eliminar un objeto o atributo del contexto cuando existen valores ausentes.

Estableceremos a continuación implicaciones entre intervalos que nos permitirán reemplazar los valores ausentes y acotar, en la medida de lo posible, el grado de desconocimiento del contexto.

Para concluir, ilustraremos estos resultados con un ejemplo.

Palabras Clave: Adquisición de conocimiento, minería de datos, contextos L-Fuzzy, conjuntos frecuentes, reglas de asociación, implicaciones entre atributos.

1. Introducción. Contextos L-Fuzzy intervalo-valorados

La necesidad de trabajar con contextos L-Fuzzy intervalo-valorados nos surge como respuesta al problema de la extracción de conocimiento de un contexto L-Fuzzy con información incompleta. Nuestra propuesta para resolver este problema pasa por transformar el contexto L-Fuzzy en un contexto L-Fuzzy intervalo-valorado, en el cual las funciones de pertenencia de los objetos y de los atributos así como la relación definida entre ellos toman valores en el conjunto de intervalos cerrados en $[0, 1]$.

El estudio de los contextos L-fuzzy intervalo-valorados ya ha sido abordado anteriormente por Burusco y Fuentes-González [4, 5] utilizando para ello los *multiconjuntos* y la *teoría de expertones*. Nosotros vamos a enfocar aquí el estudio de estos contextos desde un punto de vista diferente. Para ello, consideraremos la noción de concepto L-fuzzy definido a partir de operadores implicación [3] y lo extenderemos al caso intervalo-valorado.

Definición 1 Sea $(L, \leq)$ un retículo completo dotado de una complementación $'$ y una t-conorma S y sea $\mathcal{J}[L]$ el conjunto de intervalos de L. Sean X e Y dos conjuntos finitos(de *objetos* y *atributos* respectivamente) y R una relación L-fuzzy intervalo-valorada entre X e Y.
Llamaremos *contexto L-fuzzy intervalo-valorado* a la tupla $(\mathcal{J}[L], X, Y, R)$.

Denotaremos por $\mathcal{J}[L]^X$ y $\mathcal{J}[L]^Y$ las clases de los conjuntos L-fuzzy intervalo-valorados asociados a X y a Y respectivamente.

Definición 2 Sea $\boldsymbol{I}$ una implicación fuzzy intervalo-valorada definida en el retículo de

*Este trabajo ha sido financiado con el proyecto TIC 2003-08693 del Ministerio de Ciencia y Tecnología.

Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA2005, pp.333-340
ISBN: 84-9732-449-8

intervalos $(\mathcal{J}[L], \leq)$. Dados $A \in \mathcal{J}[L]^X$ y $B \in \mathcal{J}[L]^Y$ dos conjuntos L-fuzzy intervalo-valorados, y $R \in \mathcal{J}[L]^{X \times Y}$ una relación L-Fuzzy intervalo-valorada, definimos los *conjuntos derivados* de A y B denotados por $A_1 \in \mathcal{J}[L]^Y$ y $B_2 \in \mathcal{J}[L]^X$ respectivamente, de la siguiente manera:

$$A_1(y) = \inf_{x \in X} \{\boldsymbol{I}(A(x), R(x,y))\}$$

$$B_2(x) = \inf_{y \in Y} \{\boldsymbol{I}(B(y), R(x,y))\}$$

Llamaremos a los operadores denotados por los subíndices 1 y 2 así definidos *operadores derivación.*

Definición 3 Si representamos como A_{12} y B_{21} los conjuntos L-fuzzy intervalo-valorados $(A_1)_2$ y $(B_2)_1$ respectivamente, se definen los *operadores constructores* $\boldsymbol{\varphi}$ y $\boldsymbol{\psi}$ como:

$$\boldsymbol{\varphi}\colon \mathcal{J}[L]^X \longrightarrow \mathcal{J}[L]^X \quad /\boldsymbol{\varphi}(A) = A_{12}$$

$$\boldsymbol{\psi}\colon \mathcal{J}[L]^Y \longrightarrow \mathcal{J}[L]^Y \quad /\boldsymbol{\psi}(B) = B_{21}$$

Estos operadores constructores nos permitirán definir los conceptos L-fuzzy intervalo-valorados.

Definición 4 Si $M \in fix(\boldsymbol{\varphi})$, entonces (M, M_1) es un *concepto L-fuzzy intervalo-valorado* del contexto L-fuzzy intervalo-valorado $(\mathcal{J}[L], X, Y, R)$.

Del mismo modo que se muestra para los conceptos L-fuzzy [2], se puede demostrar de manera inmediata que también en el caso de los *conceptos L-fuzzy intervalo-valorados* la definición se podría dar equivalentemente utilizando el conjunto de puntos fijos del operador $\boldsymbol{\psi}$.

Representaremos el conjunto de conceptos L-fuzzy intervalo-valorados como $\mathcal{L}(\mathcal{J}[L], X, Y, R)$ o simplemente como $\boldsymbol{\mathcal{L}}$.

El conjunto $\mathcal{L}(\mathcal{J}[L], X, Y, R)$ dotado de la relación de orden $\preceq$ definida por $(A, B) \preceq (C, D) \Leftrightarrow A \leq C$ (o equivalentemente $(A, B) \preceq (C, D) \Leftrightarrow B \geq D$) es un retículo completo.

2. Contextos con valores ausentes

A la hora de estudiar un contexto en el que hay valores de la relación $R(x, y)$ que son desconocidos, podemos pensar en tres maneras de tratar esos *valores ausentes*, utilizando las técnicas de limpieza y transformación que se establecen en minería de datos [7]:

Ignorarlos. La opción más aconsejable sería trabajar únicamente con los datos de cuyos valores disponemos, sin embargo, en el caso del cálculo de los conceptos asociados al contexto con el que estamos trabajando debemos utilizar todos los valores de la relación $R(x, y)$, por lo que no podemos ignorar ninguno de ellos.

Eliminarlos. Para tratar el problema de los valores ausentes debemos eliminar de la relación la fila o columna correspondiente, con lo que además perderemos información. Eliminar los valores ausentes no siempre es, por tanto, una opción aconsejable y únicamente lo efectuaremos cuando estemos en uno de los siguientes casos:

- Cuando el número de valores ausentes en una fila o en una columna de la relación sea mayor que un valor máximo fijo, eliminaremos la fila (el objeto) o la columna (el atributo) correspondiente.
- Eliminaremos el objeto o el atributo para el que tenemos valores ausentes cuando no sea un objeto o un atributo *frecuente.*
- Podemos utilizar técnicas de *clustering* para clasificar los datos (tanto los objetos como los atributos) y eliminar aquellos que queden aislados.

Reemplazarlos. Por último, podemos reemplazar los valores ausentes por otros valores con los que sea posible realizar cálculos. Sin embargo, debemos tener en cuenta que al reemplazar los valores ausentes por otros estamos introduciendo información que no es real. Deberemos estudiar la manera de elegir en cada caso el valor con el que reemplazar un valor ausente, para ello:

- Elegiremos un valor con el cual representar el desconocimiento.
- Utilizaremos implicaciones entre atributos para intentar predecir el valor ausente.

3. Eliminación de valores ausentes

Para definir los objetos y atributos frecuentes vamos a utilizar la idea de *conjuntos frecuentes* [1, 7] que aparece en minería de datos.

Dado un umbral al que llamaremos soporte mínimo, un conjunto se llama *frecuente* si su soporte es mayor o igual que dicho soporte mínimo, entendiendo por soporte el cociente entre en número de veces que aparecen los elementos del conjunto dividido por el número máximo posible.

Extendiendo esta idea a la teoría de conceptos, dado un contexto formal (G, M, I) se define un atributo frecuente de la siguiente manera:

Definición 5 (Stumme et al. [8]) Sean $B \subseteq M$ y sopmin $\in [0, 1]$. Se llama *soporte* de B en el contexto al valor (perteneciente a [0,1])

$$\text{sop}(B) = \frac{|B^*|}{|G|}$$

siendo B^* el conjunto derivado de B. Se dice que B es un conjunto de atributos *frecuente* si $\text{sop}(B) \geq \text{sopmin}$.

Nosotros hemos adaptado esta definición al caso de contextos L-fuzzy y contextos L-fuzzy intervalo-valorados con el fin de decidir en qué casos podemos eliminar un valor ausente.

A partir de este punto utilizaremos $L = \mathcal{C} = \{0, 0.1, \ldots 0.9, 1\}$ y representaremos los valores ausentes de R por elementos de $\mathcal{J}[C]$.

Definición 6 [6] Sea (L, X, Y, R) un contexto L-fuzzy. Dado $A \in L^X$, definimos el *soporte de* A como el cociente (perteneciente a [0,1])

$$\text{sop}(A) = \frac{\sum_{y \in Y} A_1(y)}{|Y|}$$

De la misma manera, dado $B \in L^Y$, definimos el *soporte de* B como

$$\text{sop}(B) = \frac{\sum_{x \in X} B_2(x)}{|X|}$$

El conjunto L-fuzzy A (análogamente B) será un conjunto *frecuente* cuando dado un soporte mínimo sopmin $\in [0, 1]$, el soporte de A sea mayor o igual que sopmin (análogamente con B).

Definición 7 Sea $(\mathcal{J}[L], X, Y, R)$ un contexto L-fuzzy intervalo-valorado. Dado un conjunto L-fuzzy intervalo-valorado $A \in \mathcal{J}[L]^X$, definimos el *soporte* de A como el intervalo (perteneciente a $\mathcal{J}[0, 1]$)

$$\text{sop}(A) = \left[\frac{\sum_{y \in Y} l_{A_1}(y)}{|Y|}, \frac{\sum_{y \in Y} u_{A_1}(y)}{|Y|} \right]$$

donde la función de pertenencia del conjunto derivado A_1 viene dada por

$$A_1(y) = [l_{A_1}(y), u_{A_1}(y)]$$

Así mismo, dado un conjunto L-fuzzy intervalo-valorado $B \in \mathcal{J}[L]^Y$, definimos

$$\text{sop}(B) = \left[\frac{\sum_{x \in X} l_{B_2}(x)}{|X|}, \frac{\sum_{x \in X} u_{B_2}(x)}{|X|} \right]$$

Definición 8 Dado un conjunto L-fuzzy intervalo-valorado $A \in \mathcal{J}[L]^X$ (o análogamente $B \in \mathcal{J}[L]^Y$), y dado un soporte mínimo sopmin $\in [0, 1]$, diremos que A es un conjunto *frecuente* si se cumple que

$$[\text{sopmin}, \text{sopmin}] \leq \text{sop}(A)$$

Definición 9 Diremos que un objeto (o un atributo) de un contexto L-fuzzy intervalo-valorado es *frecuente* cuando lo es el conjunto L-fuzzy intervalo-valorado que lo representa.

Tal y como indicábamos anteriormente, cuando alguno de los valores de la relación

$R(x, y)$ es desconocido, eliminaremos el atributo o el objeto correspondiente si no es frecuente.

Sin embargo, el cálculo del soporte de cualquier conjunto precisa del conocimiento de la relación, por lo que la existencia de valores desconocidos nos impedirá conocer el soporte de los objetos y atributos del contexto. Para decidir si eliminamos un objeto o un atributo para el que tenemos un valor ausente, sustituiremos el valor ausente por $[1, 1]$ y calcularemos el soporte con la nueva tabla, lo que nos dará un valor mayor o igual que el soporte real. Eliminaremos el objeto o el atributo correspondiente si el valor obtenido es menor que el soporte mínimo fijado.

Ejemplo.

Supongamos que queremos hacer un estudio de los ingredientes de un cierto producto alimenticio para las principales marcas de dicho producto que se encuentran en el mercado. Para ello disponemos, para cada uno de los ingredientes, de las cantidades aproximadas que declaran los fabricantes por cada 100 gramos de producto.

	$I1$	$I2$	$I3$	$I4$
$M1$	[20,25]	2	10	[10,15]
$M2$	20		5	15
$M3$	25	12	10	10
$M4$	22	10	[2,5]	
$M5$	20	5	[5,10]	5

Cuadro 1: Cantidad del ingrediente Ii en la marca Mj

Consideraremos un contexto L-fuzzy intervalo-valorado $(\mathcal{J}[L], X, Y, R)$ con $L = \mathcal{C}$, el conjunto de objetos X, el conjunto de las distintas marcas y el conjunto de atributos Y, el conjunto de ingredientes. La relación R será la obtenida al transformar la tabla anterior en una tabla de intervalos cerrados y dividir cada uno de los elementos por el mayor de los elementos que aparecen en la tabla, realizando los correspondientes redondeos para que los valores pertenezcan a L. Podemos ver la relación L-fuzzy intervalo-valorada resultante en el Cuadro 2.

R	$I1$	$I2$	$I3$	$I4$
$M1$	[0. 8, 1]	[0. 1, 0. 1]	[0. 4, 0. 4]	[0. 4, 0. 6]
$M2$	[0. 8, 0. 8]		[0. 2, 0. 2]	[0. 6, 0. 6]
$M3$	[1,1]	[0,0]	[0. 4, 0. 4]	[0. 4, 0. 4]
$M4$	[0. 9, 0. 9]	[0. 2, 0. 2]	[0. 1, 0. 2]	
$M5$	[0. 8, 0. 8]	[0, 0. 2]	[0. 2, 0. 4]	[0. 2, 0. 2]

Cuadro 2: Valores de la relación R

El cálculo del retículo de conceptos no es posible en este caso debido a la existencia de dos valores desconocidos en la relación, por lo que estudiaremos en primer lugar si es posible eliminar la fila o columna correspondiente a dichos valores desconocidos.

Vamos a considerar, por ejemplo, como soporte mínimo sopmin $= 0.5$ y eliminaremos el objeto o atributo correspondiente a un valor desconocido si no es frecuente, esto es, si su soporte no es mayor o igual que dicho soporte mínimo. Para ello sustituimos los valores ausentes por el valor $[1, 1]$ y calculamos el soporte de los objetos $M2$ y $M4$ y de los atributos $I2$ e $I4$ con la nueva relación $\widehat{R}$.

$\widehat{R}$	$I1$	$I2$	$I3$	$I4$
$M1$	[0. 8, 1]	[0. 1, 0. 1]	[0. 4, 0. 4]	[0. 4, 0. 6]
$M2$	[0. 8, 0. 8]	[1, 1]	[0. 2, 0. 2]	[0. 6, 0. 6]
$M3$	[1,1]	[0,0]	[0. 4, 0. 4]	[0. 4, 0. 4]
$M4$	[0. 9, 0. 9]	[0. 2, 0. 2]	[0. 1, 0. 2]	[1, 1]
$M5$	[0. 8, 0. 8]	[0, 0. 2]	[0. 2, 0. 4]	[0. 2, 0. 2]

Cuadro 3: Valores de la relación $\widehat{R}$, sin valores ausentes

El conjunto L-fuzzy intervalo-valorado asociado al objeto $M2$ es

$$M2 = \{M1/[0,0], M2/[1,1], M3/[0,0], M4/[0,0], M5/[0,0]\}$$

Para calcular su conjunto derivado hemos considerado la R-implicación de Lukasiewicz y hemos obtenido:

$$(M2)_1 = \{I1/[0.8, 0.8], I2/[1,1], I3/[0.2, 0.2], I4/[0.6, 0.6]\}$$

y por lo tanto su soporte será

$$\text{sop}(M2) = \left[\frac{2.6}{4}, \frac{2.6}{4}\right]$$

Con el soporte mínimo que hemos considerado sopmin $= 0.5$, $M2$ será un objeto *frecuente*.

Análogamente podemos calcular el soporte de los conjuntos intervalo-valorados asociados a la marca $M4$ y a los ingredientes $I2$ e $I4$ y los valores que se obtienen son los siguientes:

$$\text{sop}(M4) = \left[\frac{2.2}{4}, \frac{2.3}{4}\right]$$
$$\text{sop}(I2) = \left[\frac{1.3}{5}, \frac{1.5}{5}\right]$$
$$\text{sop}(I4) = \left[\frac{2.6}{5}, \frac{2.8}{5}\right]$$

Se observa que, con el soporte mínimo que estamos considerando, el único conjunto que no es frecuente es el asociado a $I2$, por lo que sería éste el único que eliminaríamos.

Podemos eliminar un mayor número de objetos o de atributos si elegimos un valor más alto para el soporte mínimo. Si en el caso anterior consideramos como soporte mínimo sopmin $= 0.55$, además de $I2$ eliminaremos también $I4$. Sin embargo, debemos tener en cuenta que la eliminación de objetos o de atributos del contexto implica una pérdida de información, por lo que no será aconsejable trabajar con soportes mínimos con valores muy altos.

4. Reemplazamiento de valores ausentes. Implicaciones entre atributos

Una vez eliminados todos los objetos y atributos con valores ausentes que no sean frecuentes, nuestra siguiente tarea será elegir la manera de representar esos elementos desconocidos por medio de un valor que nos permita trabajar con ellos.

Podemos considerar que el *desconocimiento* es equiparable a la imprecisión total, y, por lo tanto, representarlo mediante el intervalo $[0, 1]$.

La utilización del intervalo $[0, 1]$ para representar el desconocimiento nos va a proporcionar en general valores con una amplitud de intervalo grande, ya que cuando la información de que disponemos no es completa el conocimiento que podemos extraer estará dotado de un cierto grado de ambigüedad. En cualquier caso, intentaremos reducir al máximo esta ambigüedad mediante implicaciones entre atributos.

Con el fin de establecer implicaciones entre atributos vamos a utilizar una importante herramienta para la extracción de conocimiento como son las *reglas de asociación*.

La idea de regla de asociación fue presentada por primera vez por Agrawal et al. en [1] como una implicación de la forma $B \Rightarrow C$, donde $B, C \subseteq M$ son dos conjuntos de atributos denominados respectivamente *antecedente* y *consecuente*, que se verifica con un cierto soporte y grado de confianza.

El *soporte* de una regla $B \Rightarrow C$ es el porcentaje de veces que la regla predice correctamente, esto es, el porcentaje de veces que se tienen los atributos de B y los de C

$$\text{sop}(B \Rightarrow C) = \text{sop}(B \cup C)$$

La *confianza* de una regla es el porcentaje de veces que la regla se cumple cuando se puede aplicar, es decir, el porcentaje de veces que conteniendo los atributos de B también se tienen los de C

$$\text{conf}(B \Rightarrow C) = \frac{\text{sop}(B \cup C)}{\text{sop}(B)}$$

Utilizando las reglas de asociación y la definición de soporte de un conjunto L-fuzzy, definimos en [6] el soporte y el grado de confianza de una implicación entre atributos de la siguiente forma.

Definición 10 Sea (L, X, Y, R) un contexto L-fuzzy y sean $B, C \subseteq L^Y$ dos conjuntos de atributos. El *soporte* de la implicación $B \Rightarrow C$ viene dado por

$$\text{sop}(B \Rightarrow C) = \text{sop}(B \cup C) = \frac{\sum_{x \in X} (B \cup C)_2(x)}{|X|}$$

y representa el porcentaje de objetos que comparten los atributos de B y de C.
El grado de *confianza* de la implicación es

$$\text{conf}(B \Rightarrow C) = \frac{\text{sop}(B \cup C)}{\text{sop}(B)} = \frac{\sum_{x \in X} (B \cup C)_2(x)}{\sum_{x \in X} B_2(x)}$$

y representa el porcentaje de objetos que verifican la implicación, es decir, el porcentaje de objetos que teniendo los atributos de B en un cierto grado tienen también los de C en el mismo grado.

Utilizando la definición de soporte de un conjunto L-fuzzy intervalo-valorado, podemos extender la definición de soporte y confianza de una implicación entre atributos al caso intervalo-valorado.

Definición 11 Sea $(\mathcal{J}[L], X, Y, R)$ un contexto L-fuzzy intervalo-valorado. Dados los conjuntos de atributos $B, C \in \mathcal{J}[L]^Y$, definimos el *soporte* de la implicación $B \Rightarrow C$, $\text{sop}(B \Rightarrow C)$, como el intervalo

$$\left[\frac{\sum_{x\in X} l_{(B\cup C)_2}(x)}{|X|}, \frac{\sum_{x\in X} u_{(B\cup C)_2}(x)}{|X|}\right]$$

donde la función de pertenencia del conjunto derivado B_2 viene dada por

$$B_2(x) = [l_{B_2}(x), u_{B_2}(x)]$$

La confianza de la implicación, $\text{conf}(B \Rightarrow C)$, vendrá dada por el intervalo

$$\left[\frac{\sum_{x\in X} l_{(B\cup C)_2}(x)}{\sum_{x\in X} u_{B_2}(x)}, \frac{\sum_{x\in X} u_{(B\cup C)_2}(x)}{\sum_{x\in X} l_{B_2}(x)} \wedge 1\right]$$

Vamos a utilizar las implicaciones entre atributos para intentar estimar el valor de los datos ausentes. Tal y como hemos indicado, el nivel de confianza de una implicación entre atributos nos indica el porcentaje de objetos que teniendo los atributos del antecedente en un cierto grado, poseen también los del consecuente en el mismo grado.

Ejemplo.

Si volvemos al ejemplo anterior, teníamos la relación intervalo-valorada dada por el Cuadro 4 en donde los dos valores que eran desconocidos los hemos sustituido por el valor $[0, 1]$.

Para poder estimar dichos valores desconocidos vamos a estudiar las implicaciones entre los distintos atributos con el fin de encontrar aquellas que se verifiquen con un nivel de confianza alto.

R	$I1$	$I2$	$I3$	$I4$
$M1$	$[0.8, 1]$	$[0.1, 0.1]$	$[0.4, 0.4]$	$[0.4, 0.6]$
$M2$	$[0.8, 0.8]$	$[0, 1]$	$[0.2, 0.2]$	$[0.6, 0.6]$
$M3$	[1,1]	[0,0]	$[0.4, 0.4]$	$[0.4, 0.4]$
$M4$	$[0.9, 0.9]$	$[0.2, 0.2]$	$[0.1, 0.2]$	$[0, 1]$
$M5$	$[0.8, 0.8]$	$[0, 0.2]$	$[0.2, 0.4]$	$[0.2, 0.2]$

Cuadro 4: Valores de la relación R

Estudiaremos las implicaciones entre atributos en las que aparezcan los atributos $I2$ e $I4$.

• Si consideramos, por ejemplo, los atributos $I1$ e $I2$, podemos calcular con qué soporte y con qué nivel de confianza se verifica la implicación $I_1 \Rightarrow I2$:

Los conjuntos L-fuzzy intervalo-valorados asociados a los conjuntos de atributos $I1$ e $I1 \cup I2$ serán respectivamente

$$I1 = \{I1/[1,1], I2/[0,0], I3/[0,0], I4/[0,0]\}$$

$$(I1 \cup I2) = \{I1/[1,1], I2/[1,1], I3/[0,0], I4/[0,0]\}$$

Calculamos sus conjuntos derivados utilizando, por ejemplo, la R-implicación de Lukasiewicz

$$(I1)_2 = \{M1/[0.8,1], M2/[0.8,0.8], M3/[1,1], M4/[0.9,0.9], M5/[0.8,0.8]\}$$

$$(I1 \cup I2\ )_2 = \{M1/[0.1,0.1], M2/[0,0.8], M3/[0,0], M4/[0.2,0.2], M5/[0,0.2]\}$$

y por lo tanto el soporte y el nivel de confianza de la implicación serán

$$\text{sop}(I1 \Rightarrow I2) = \left[\frac{0.3}{5}, \frac{1.3}{5}\right] = [0.06, 0.26]$$

$$\text{conf}(I1 \Rightarrow I2) = \left[\frac{0.3}{4.5}, \frac{1.3}{4.3}\right] = [0.06, 0.3]$$

En este caso tanto el nivel de confianza como el soporte de la implicación son próximos a 0. Podemos deducir que entre un 6 %

y un 26 % de las marcas contienen una gran cantidad de los ingredientes $I1$ e $I2$, y que únicamente entre un 6 % y un 30 % de las marcas que tienen el ingrediente $I1$, también contienen al menos la misma cantidad del ingrediente $I2$. Se tiene pues una implicación que se verifica con un nivel de confianza bajo y que, por lo tanto, no nos proporciona información sobre el grado de pertenencia del ingrediente $I2$ a la marca $M2$.

• Si consideramos la implicación $I3 \Rightarrow I4$, obtendremos un soporte y un nivel de confianza dados por:

$$\text{sop}(I3 \Rightarrow I4) = \left[\frac{1.2}{5}, \frac{1.4}{5}\right] = [0.24, 0.28]$$

$$\text{conf}(I3 \Rightarrow I4) = \left[\frac{1.2}{1.6}, 1\right] = [0.75, 1]$$

Entre un 24 y un 28 % de las marcas utilizan una gran cantidad de los ingredientes $I3$ e $I4$. El nivel de confianza con que se verifica la implicación es muy alto, como mínimo en un 75 % de las marcas en las que aparece el ingrediente $I3$ tendremos al menos la misma cantidad del ingrediente $I4$. Esto nos permite extender la regla al caso de la marca $M4$ para la cual desconocemos la cantidad del ingrediente $I4$, y supondremos que también en este caso se tiene como mínimo la misma cantidad que del ingrediente $I3$, esto es

$$R(M4, I4) \geq R(M4, I3) = [0.1, 0.2]$$

Obtenemos una condición que nos permite reducir el desconocimiento, y representar ese valor con un intervalo de amplitud un poco menor, $R(M4, I4) = [0.1, 1]$.

• El soporte y el nivel de confianza con que se cumple la implicación $I4 \Rightarrow I1$ son:

$$\text{sop}(I4 \Rightarrow I1) = \left[\frac{1.6}{5}, \frac{2.7}{5}\right] = [0.32, 0.54]$$

$$\text{conf}(I4 \Rightarrow I1) = \left[\frac{1.6}{2.8}, 1\right] = [0.57, 1]$$

de donde podemos concluir que el porcentaje de marcas en las que aparece el ingrediente $I1$ al menos en la misma proporción en que aparece el ingrediente $I4$ es como mínimo de un 57 %. Tenemos pues un valor para el nivel de confianza del cumplimiento de la implicación bastante alto, y por lo tanto la probabilidad de que en la marca $M4$ también se cumpla será bastante alta. Supondremos entonces que

$$R(M4, I4) \leq R(M4, I1) = [0.9, 0.9]$$

y asignaríamos el valor $R(M4, I4) = [0, 0.9]$

Podemos conseguir una imprecisión menor si consideramos las dos implicaciones, de esa forma:

$$[0.1, 0.2] \leq R(M4, I4) \leq [0.9, 0.9]$$

y para que esto se cumpla, la imprecisión será menor o igual que las que nos da el valor $R(M4, I4) = [0.1, 0.9]$.

De esta manera podríamos analizar todas las implicaciones entre atributos con algún valor desconocido, con el fin de conseguir relaciones entre ellos que nos permitan acotar el grado de imprecisión de cada uno de estos valores.

Observaciones:

1. El nivel de confianza de la implicación nos proporciona un porcentaje de veces en que ésta se cumple cuando se da el antecedente, por lo que cuanto menor sea esta confianza mayor será la probabilidad que tenemos de cometer errores al utilizar dicha implicación con el fin de estimar valores desconocidos. Nos interesará por lo tanto trabajar con implicaciones cuyo nivel de confianza sea alto.

2. El valor $[0, 1]$ que utilizamos para representar los valores desconocidos hará que el nivel de confianza de una implicación en la que aparece el correspondiente argumento disminuya, por lo que será muy difícil conseguir niveles de confianza cercanos al 100 %.

5. Líneas futuras

En trabajos futuros intentaremos mejorar el proceso de eliminación de valores ausentes (por ejemplo, mediante técnicas de fuzzy clustering) y, tras proceder a la construcción del retículo de conceptos L-Fuzzy, analizaremos qué influencia tienen esos valores ausentes en los resultados obtenidos.

También será interesante utilizar minería de reglas de asociación para quedarnos con aquellas implicaciones cuyos soportes y confianzas queden por encima de unos determinados umbrales.

Referencias

[1] R.Agrawal, T.Imielinski, A.Swami. Mining association rules between sets of items in large databases. *Proceedings of the ACM SIGMOD International Conference on Management of Data.* Washington, (1993), pp.207-216.

[2] A.Burusco, R.Fuentes-González. The Study of the L-Fuzzy Concept Lattice. *Mathware and Computing.* 1 No.3 (1994), pp.209-218.

[3] A.Burusco, R.Fuentes-González. Concept lattices defined from implication operators. *Fuzzy Sets and Systems.* 114 No.1 (2000), pp.431-436.

[4] A.Burusco, R.Fuentes-González. The study of the interval-valued contexts. *Fuzzy Sets and Systems.* 121 No.3 (2001), pp.69-82.

[5] A.Burusco, R.Fuentes-González. Contexts with multiple weighted values. *International Journal of Uncertanty, Fuzziness and Knowledge-Based Systems.* 9 No.3 (2001), pp.355-368.

[6] A.Burusco, R.Fuentes-González, C.Alcalde. Utilización de reglas de asociación en la teoría de conceptos L-Fuzzy. *ESTYLF 2004.* Jaén (2004), pp.303-308.

[7] J.Hernández, M.J.Ramírez, C.Ferri. *Introducción a la minería de datos.* Pearson Prentice Hall, Madrid, (2004).

[8] G.Stumme, R.Taouil, Y.Bastide, N.Pasquier, L.Lakhal. Computing iceberg concept lattices with TITANIC. *Data & Knowledge Engineering.* 42 (2002), pp.189-222.