Extending Knowledge Based Redundancy in Association Rules with Imprecise Knowledge

J. Díaz, G. Negrín, C. Molina, and M. Vila

***Abstract*—Association Rules Mining is one of the most studied and widely applied fields in Data Mining. However, the discovery models usually result in a very large set of rules; so the analysis capability, from the user point of view, is diminishing. Hence, it is difficult to use the found model in order to assist decision-making process. The previous handicap is heightened in presence of redundant rules in the final set. In this work we study a way to eliminate redundancy in association rules with uncertainty, with imprecise user prior knowledge. A post-processing method is developed to eliminate this kind of redundancy, using association rules known by the user. Our proposal allows to find more compact models of association rules to ease its use in the decision-making process. The developed experiments have shown that the reduction using certainty factor has a slightly better behavior. The most important contribution of this paper is the definition of a mechanism to remove knowledge based redundancy using Dempster-Schaffer Theory and the Certainty Factor Model.**

***Index Terms*—Imprecise Association Rule Mining, Redundant Rules, Knowledge guided post-processing, Dempster-Schaffer Theory, Certainty Factor Model.**

I. Introducción

La MINERÍA de reglas de asociación [1] es uno de los campos más estudiados en la minería de datos. Las reglas de asociación se encargan del descubrimiento de relaciones entre atributos en grandes bases de datos. Una regla de asociación se presenta como una implicación de la forma $X \rightarrow Y$ donde X e Y son conjuntos de ítems que de manera general cumplen la propiedad $X \cap Y = \emptyset$, pero no es obligatorio. Estas reglas son del tipo *si-entonces*, por lo tanto, son fáciles de entender por agentes humanos y también de utilizar por parte de los últimos en el proceso de toma de decisiones.

Existen un gran número de medidas para evaluar la relevancia de una regla de asociación [5], pero, no hay un consenso en la comunidad científica de cuáles son las mejores [6]. El marco más conocido es el del soporte y la confianza.

El soporte de una regla se denota $supp(X \rightarrow Y)$ y se refiere a la porción de la transacción de la base de datos para la cual $X \cup Y$es verdadera. Representa la frecuencia de aparición del itemset en una base de datos, por lo tanto, brinda una idea de cuán frecuente es un patrón en un dataset.

La confianza de una regla de asociación se denota $conf(X \rightarrow Y)$, es una medida para conocer la probabilidad condicional de la ocurrencia de Y en las transacciones de la base de datos que contienen a X. Representa cuánta influencia tiene la ocurrencia de X en la ocurrencia de Y.

Cada regla de asociación extraída debe superar el umbral de soporte y confianza para ser interesante. Las reglas de asociación con altas frecuencias son usualmente conocidas por el usuario, por lo tanto, para encontrar las reglas interesantes sin pérdida de información, deben definirse valores bajos de soporte y confianza. Con los valores bajos de estas medidas se presenta un gran número de reglas extraídas, incluso para datasets pequeños [9]. Este gran número de reglas extraídas hace el modelo no adecuado para su análisis por parte del usuario.

Una gran parte de las reglas presentadas al usuario son irrelevantes porque son obvias, muy generales, demasiado específicas o simplemente porque no ofrecen información nueva.

Las reglas que no brindan nueva información se denominan redundantes y tienen un gran impacto en el número total de reglas en el modelo. Recientemente se ha propuesto un enfoque basado en conocimiento previo para tratar el problema de la redundancia [8], en el que se define la redundancia basada en el conocimiento previo y un algoritmo para eliminarla. Este enfoque aprovecha la habilidad de un experto para expresar conocimiento del dominio usando reglas *si-entonces*, siguiendo la idea que, si una regla incluye algunos elementos conocidos por el usuario, entonces, puede ser eliminada sin pérdida de información. Sin embargo, no toma en cuenta la naturaleza imprecisa del conocimiento humano y por lo tanto limita la capacidad de expresar su conocimiento al usuario.

Este artículo se enfoca en generalizar el método de eliminación de redundancia basada en conocimiento previo [8] para trabajar con incertidumbre. Esta generalización hace posible construir modelos que están más cerca del usuario y representan de una manera más precisa el dominio.

TABLA I
Ejemplo De Transacciones En Un Dominio

Cliente	*Ingreso*	*Balance*	*Sexo*	*Sin empleo*	*Préstamo*
C1	Alto	Alto	F	No	Sí
C2	Alto	Alto	M	No	Sí
C3	Bajo	Bajo	M	No	No
C4	Bajo	Alto	F	Sí	Sí
C5	Bajo	Alto	M	Sí	Sí
C6	Bajo	Bajo	F	Sí	No
C7	Alto	Bajo	M	No	Sí
C8	Alto	Bajo	F	Sí	Sí
C9	Bajo	Medio	M	Sí	No
C10	Alto	Medio	M	No	Sí
C11	Bajo	Medio	F	Sí	No
C12	Bajo	Medio	M	No	No

Considerando el dominio sobre el que se sustenta la Tabla 1, un especialista podría tener el conocimiento siguiente: si los clientes tienen un balance alto entonces pagan los préstamos. El modelo de reducción de redundancia planteado en [8] aprovecha ese conocimiento para eliminar las reglas que no aporten nueva información. Sin embargo, es más frecuente que los especialistas incorporen determinado grado de

incertidumbre dentro de su conocimiento del dominio, de esta forma el especialista expresaría la regla anterior como el 80% de los clientes que tienen un balance alto pagan sus préstamos. La propuesta realizada en [8] no es capaz de gestionar este tipo de reglas ya que considera todo lo expresado en el conocimiento como certeza absoluta. Este modelo se aleja de la realidad y de la manera en que gestionan el conocimiento los agentes humanos. Es por ello que este trabajo se pretende extender el modelo presentado en [8] añadiéndole la capacidad de utilizar conocimiento impreciso para la definición del conocimiento del experto.

El artículo se organiza de la manera siguiente: La sección 2 aborda la redundancia del conocimiento previo. En la sección 3 se realiza la formalización matemática del problema. En la sección 4 se trata el tema de la introducción de la incertidumbre en el conocimiento previo. En la sección 5 se abordan los experimentos y la discusión de los resultados.

II. Redundancia Basada En El Conocimiento

Una de las definiciones más estudiadas para la redundancia en reglas de asociación es la propuesta por Bastide [4] "una regla de asociación es redundante si contiene la misma información o información menos general que la información contenida en otra regla de la misma utilidad y relevancia". La mayoría de las investigaciones que tratan el problema de la redundancia utilizan una interpretación probabilística de "la misma información o información menos general", pero, de acuerdo a [3] este tipo de interpretación ha alcanzado su desempeño máximo y todavía no es suficiente para producir modelos de reglas de asociación que sean fáciles de utilizar por agentes humanos.

Definición 1. Redundancia del conocimiento previo [8]: Sea S un conjunto de reglas de asociación y S_c un conjunto de reglas previamente conocidas, definido sobre el mismo dominio de S. Una regla de asociación $R: X \rightarrow Y \in S$ es redundante respecto a S_c si hay una regla $R': X' \rightarrow Y' \in S_c$ y que cumpla alguna de las siguientes condiciones.

1. $X' \subseteq X \wedge Y' \cap Y \neq \{\emptyset\}$
2. $X' \subseteq X \wedge \exists R'': X'' \rightarrow Y'' \in S_c: X'' \subseteq Y' \wedge Y \subseteq Y''$
3. $X' \subseteq X \wedge Y' \cap X \neq \{\emptyset\}$
4. $X' \subseteq Y \wedge Y' \cap Y \neq \{\emptyset\}$

Para detectar redundancia en modelos de reglas de asociación se propuso el Algoritmo 1.

Entrada: Un conjunto de reglas conocidas S_c, una regla $R_i: X \rightarrow Y$
Salida: Valor boolean verdadero si la regla R_i es redundante

1: $i = 0$
2: $n = |Y|$
3: **while** $i < n$ **do**
4: **if** $Y[i] \in X^{+}_{Sc \cup X \rightarrow (Y - \{Y[i]\})}$ **then**
5: **return** $true$
6: **end**
7: $i = i + 1$
8: **end**
9: $i = 0$
10: $n = |X|$
11: **while** $i < n$ **do**
12: **if** $X[i] \in (X - X[i])^{+}_{Sc \cup (X - X[i]) \rightarrow Y}$ **then**
13: **return** $true$
14: **end**
15: $i = i + 1$
16: **end**
17: **return** $false$

Fig. 1. Algoritmo de detección de la redundancia basada en el conocimiento previo.

Las entradas del algoritmo 1 son un conjunto de reglas S_c y una regla R_i en la forma $X \rightarrow Y$, la última es el objeto de estudio. Se chequean los ítems en R_i en busca de redundancia en dos estructuras repetitivas. La primera (líneas de la 3 a la 8) itera sobre los ítems en Y usando el algoritmo del cierre [12] para verificar si el elemento es redundante (línea 4), si el ítem es redundante se devuelve $true$ (línea 5), si no es así, se comprueba el siguiente ítem (línea 7). El segundo ciclo (líneas de la 11 a la 16) itera de la misma manera pero sobre los ítems de X. Si después de la ejecución de ambos ciclos no hay elementos redundantes se retorna $false$ (línea 17) y termina la ejecución.

El dominio de conocimiento consiste en relaciones entre atributos que se conocen previamente por el usuario. Son resultado de la experiencia del experto en el área de trabajo. Por lo tanto, este conocimiento es considerado de alguna manera más comprensible que las reglas extraídas de un dataset en particular que contendría información parcial. El usuario puede representar el conocimiento previo en diferentes vías tal como redes semánticas, ontologías, entre otros.

El conocimiento del usuario es incorporado al modelo usando el formato de reglas de asociación y la razón para esto es que el experto está interesado en el descubrimiento de reglas de asociación, pero, frecuentemente el usuario tiene cierta incertidumbre acerca del conocimiento. Por ejemplo, en el análisis de la canasta de compra, la asociación entre mantequilla y pan se expresa preferentemente con un grado de incertidumbre similar a: el 80% de las personas que compran mantequilla también compran pan, en lugar de lo absoluto que las personas que compran mantequilla también compran pan. De acuerdo a esto las reglas de asociación presentadas en el conocimiento previo tienen la forma ($X \rightarrow Y\ certeza = valor$). Los valores de certeza deben estar definidos en el intervalo $[0; 1]$.

Es necesario incorporar un grado de certeza a las asociaciones que son parte del conocimiento del usuario para incorporar la posibilidad de expresar conocimiento en una manera más certera. Pero la incertidumbre generada produce nuevos problemas que deben ser afrontados.

III. Formalización del Problema

Sea D una base de datos, A una técnica para la minería de reglas de asociación sobre D y S_c una representación del conocimiento previo que contiene el grado de certeza para cada regla. El conjunto R contiene las reglas minadas. Un subconjunto R' de R contiene las reglas que se pueden derivar de S_c, por lo tanto son redundantes. Es necesario prestar atención a dos elementos importantes:

1. El mismo modelo de reglas R puede tener diferentes modelos redundantes R' asociados a usuarios con diferentes conocimiento previo.

2. El conocimiento de los usuarios puede ser modificado dentro del proceso, por lo tanto, la determinación de las reglas redundantes es un proceso dinámico e interactivo.

El conjunto de reglas potencialmente interesantes es $R - R'$. Es usual que este resultado sea considerablemente menor que R, por lo tanto, es deseable mostrar solo esas reglas.

El problema de eliminar redundancia basada en el conocimiento de un conjunto de reglas de asociación se define como: dado un conjunto de reglas de asociación R y conocimiento previo del usuario S_c encontrar el conjunto de reglas no redundantes $R - R'$ en un momento dado.

La introducción de incertidumbre al modelo de conocimiento hace necesario tratar con dos nuevos problemas:

1. ¿Cómo propagar la incertidumbre hacia las reglas derivadas?
2. ¿Cuándo un grado de incertidumbre hace a una regla no redundante?

IV. Redundancia Basada En El Conocimiento Con Incertidumbre

En la bibliografía se estudian algunos enfoques para tratar la incertidumbre en Inteligencia Artificial, para tener una perspectiva histórica ver [11]

Dos de los más simples y más usadas alternativas son la Teoría de la Evidencia de Dempster-Shaffer [7], y el Factor de Certeza [10]. Ambos pueden ser usados para construir un modelo de propagación de la incertidumbre para redundancia basada en el conocimiento con incertidumbre.

A. Teoría de la Evidencia

La teoría de la evidencia de Dempster-Schaffer (DS por sus siglas en inglés) asume un marco finito de discernimiento $\theta = \{\theta_1, \dots, \theta_n\}$ donde los elementos θ_i representan la hipótesis y son exhaustivos y exclusivos. Una función de asignación básica de probabilidad μ que vincula un elemento en θ con el valor de certeza que una evidencia provee a la hipótesis.

La redundancia basada en el conocimiento previo con conocimiento impreciso se modela usando DS de la manera siguiente:

- Hay dos hipótesis; 1- la regla es redundante y 2- la regla no es redundante.
- Cada condición verdadera en la definición 1 genera una nueva evidencia.
- La *creencia* definida por el usuario se usa como evidencia para la hipótesis de "la regla es redundante" y para la hipótesis de "la regla no es redundante"

La combinación de evidencia se lleva a cabo usando la regla de Dempster de la combinación, ecuación 1.

$$\mu_1 \oplus \mu_2(x) = \begin{cases} 0 & si\ x = \emptyset \\ \dfrac{\sum_{S_1 \cap S_2 = x} \mu_1(S_2) \times \mu_2(S_2)}{1 - \sum_{S_1 \cap S_2 = \emptyset} \mu_1(S_2) \times \mu_2(S_2)} & en\ otro\ caso \end{cases}$$

Ecuación 1.

Para presentar un ejemplo de aplicación, supóngase que el conocimiento previo de un usuario $S_c = \{A \to C\ certeza = 0.8, B \to C\ certeza = 0.6\}$ y la regla $R = A \to BC$. Si se prueba a R para detectar redundancia basada en el conocimiento previo con incertidumbre usando el modelo DS se tiene:

- Dos hipótesis 1- $A \to BC$ es redundante (R) y 2- $A \to BC$ no es redundante $(\neg R)$.
- Dos fuentes de evidencia 1- $P_{A \to C}(R) = 0.8$ y 2- $P_{B \to C}(R) = 0.6$.

Las evidencias se combinan de acuerdo a la ecuación 1 de la manera siguiente:

- $\mu_1(R) = 0.8$
- $\mu_1(\neg R) = 0.2$
- $\mu_2(R) = 0.6$
- $\mu_2(\neg R) = 0.4$
- $\mu_1 \oplus \mu_2(R) = \left\{\frac{0.8 \times 0.6}{1 - (0.8 \times 0.4 + 0.2 \times 0.6)} = 0.857\right.$
- $\mu_1 \oplus \mu_2(\neg R) = \left\{\frac{0.4 \times 0.2}{1 - (0.8 \times 0.4 + 0.2 \times 0.6)} = 0.143\right.$

Hay una particularidad importante en el modelo de DS y esta radica en que si alguna regla tiene un factor de certeza por debajo 0.5, la certeza combinada será menor que la certeza de la otra regla. Este comportamiento se acepta para tratar con evidencias con conflictos, pero este no es el caso. Por tanto, si una regla tiene una certeza menor del 0.5, la regla será descartada como evidencia.

B. Factor de Certeza

El modelo del Factor de Certeza (CF por sus siglas en inglés) es un método para tratar con la incertidumbre en sistemas basados en reglas. CF asume dos elementos:

1. El conjunto de hipótesis es exclusivo y exhaustivo
2. Todas las evidencias son condicionalmente independientes

Las reglas en CF, también, se expresan como sentencias $si - entonces$ de la forma $e \to h$ donde e denota una evidencia y h una hipótesis. La incertidumbre se modela con la adición de un valor de certeza cf. El valor de cf se restringe al intervalo $[-1; 1]$ con el significado siguiente:

- Si $-1 \leq cf \leq 0$ la creencia en la regla decrece.
- Si $0 \leq cf \leq 1$ la creencia en la regla crece.

El modelo CF incluye cuatro funciones para combinar evidencias.

1. Combinación paralela: se utiliza para combinar dos reglas A y B con la misma hipótesis.

$$CF_{AB} = \begin{cases} CF_A + CF_B - CF_A \times CF_B & si\ CF_A, CF_B \geq 0 \\ CF_A + CF_B + CF_A \times CF_B & si\ CF_A, CF_B \leq 0 \\ \dfrac{CF_A + CF_B}{1 - \min(|CF_A|, |CF_B|)} & en\ otro\ caso \end{cases}$$

Ecuación 2.

2. La combinación serie: se utiliza para combinar dos reglas A y B cuando la hipótesis de A es la evidencia de B.

$$CF_{AB} = \begin{cases} CF_A \times CF_B & si\ CF_A \geq 0 \\ 0 & si\ CF_A, CF_B \leq 0 \end{cases}$$

Ecuación 3.

3. Conjunción de evidencia: se utiliza para combinar reglas A y B cuando su conjunción es la evidencia de otra regla.

$$CF_{A\&B} = \min(CF_A, CF_B)$$
Ecuación 4.

4. Disyunción de evidencia: se utiliza para combinar reglas A y B cuando su disyunción es la evidencia de otra regla.

$$CF_{A|B} = \max(CF_A, CF_B)$$
Ecuación 5.

La redundancia basada en el conocimiento con conocimiento impreciso se modela usando CF de la manera siguiente:

- Hay dos hipótesis: 1- la regla es redundante y 2- la regla no es redundante.
- Para cada regla $R_i \in S_c$ tal que R_i cumple alguna de las condiciones en la definición 1, se crea una regla de la forma $si\ R_i\ entonces\ R\ es\ redudante$. La nueva regla tendrá un cf igual a la combinación serie definida para R_i.
- Si hay dos reglas $R_1, R_2 \in S_c$ tales que cumplen la condición 2 en la definición 1, se crea una regla de la forma $si\ R_1, R_2\ entonces\ R\ es\ redudante$. La nueva regla tendrá un cf igual a la combinación serie definida para R_1 y R_2.
- El cf para la hipótesis "la regla es redundante" se calcula por combinación paralela de la creada.

Para mostrar un ejemplo de aplicación, supóngase que se tiene un conocimiento $S_c = \{A \rightarrow C\ certeza = 0.8,\ B \rightarrow C\ certeza = 0.6\}$ y la regla $R = A \rightarrow BC$. Si se prueba a R para detectar redundancia basada en el conocimiento previo con incertidumbre usando el modelo CF se tiene:

- Dos hipótesis 1- $A \rightarrow BC$ es redundante (R) y 2- $A \rightarrow BC$ no es redundante $(\neg R)$.
- Dos reglas asociadas a condiciones en la definición 1. 1- $si\ A \rightarrow C\ entonces\ A \rightarrow BC\ es\ redundante\ cf = 0.8$ y 2- $si\ B \rightarrow C\ entonces\ A \rightarrow BC\ es\ redundante\ cf = 0.6$.
- Las evidencias se combinan usando la combinación paralela $CF_{1,2} = CF_1 + CF_2 - CF_1 \times CF_2 = 0.92$

C. Algoritmo para Detectar Redundancia Basada en el Conocimiento con Incertidumbre

Para determinar si una regla R_i tiene redundancia basada en el conocimiento previo respecto a un conocimiento impreciso S_c, se deben seguir los pasos siguientes:

- Encontrar todos los elementos redundantes en R_i. Para ello es necesario modificar el algoritmo 1.
- Encontrar las reglas S_c que contengan evidencia de los elementos redundantes en R_i. El algoritmo de cierre debe también ser modificado para cumplir este paso.
- Calcular la combinación de cada evidencia.

El algoritmo 2 muestra el pseudo-código necesario para implementar estos pasos, haciendo algunas modificaciones al algoritmo 1.

Entrada: Un conjunto de reglas con incertidumbre conocidas S_c, una regla R_i: $X \rightarrow Y$
Salida: Un valor de $cf \in [0,1]$ indicando la certeza de la redundancia de R_i

```
1:  cf = 0
2:  F = S_c ∪ R_i
3:  redundant = false
4:  rules = ∅
5:  foreach itemA ∈ X do
6:    if cf_closure(F, X − {A}, A) ≠ 0 then
7:      redundant = true
8:      rules = rules ∪ cf_closure(F, X − {A}, A)
9:    end
10: end
11: foreach itemW ∈ Y do
12:   if cf_closure(F − {R_i} ∪ X → Y − {W}, X, W) ≠ 0 then
13:     redundant = true
14:     rules = rules ∪ cf_closure(F − {R_i} ∪ X → Y − {W}, X, W)
15:   end
16: end
17: if redundant == true then
18:   cf = combine(rules, model)
19: end
20: return cf
```

Fig. 2. Algoritmo de detección de la redundancia basada en el conocimiento previo con incertidumbre.

Las entradas para el algoritmo 1 y 2 son las mismas, pero las reglas en el conjunto de reglas conocidas S_c en el algoritmo 2 tienen un grado de certeza. La salida difiere entre los algoritmos ya que el primero retorna un valor lógico mostrando la redundancia en R_i y el segundo es un grado de redundancia asociado a R_i, un valor entre $[0;1]$. Para encontrar la redundancia en R_i se usan dos estructuras repetitivas para explorar los elementos del antecedente (líneas de las 5 a la 10) y para explorar los elementos del consecuente (líneas de la 11 a la 16). El primer ciclo chequea el grado de redundancia de cada elemento en X usando el procedimiento $cf\_closure$ (línea 6), ver algoritmo 3, si algún elemento tiene un grado de redundancia por encima de cero, se marca como redundante (línea 7). Las reglas que justifican la redundancia del elemento se almacenan (línea 8) para ser combinadas más tarde. El segundo ciclo hace el mismo procedimiento, pero, sobre los elementos de Y. Una vez que se chequean los elementos de X y de Y en busca de redundancia, si alguno es redundante (línea 17), las reglas con evidencia de redundancia se combinan (línea 18). Esta combinación puede ser llevada a cabo usando el modelo CF o el DS. Finalmente, el grado de redundancia se devuelve (línea 20).

Entrada: Un conjunto de reglas F, itemset X, ítem Z
Salida: Conjunto de regla en F necesitados para incluir a Z en el cierre de X

```
1:  closure = X
2:  rules = ∅
3:  foreach itemx ∈ X do
4:    rules[item] = ∅
5:  end
6:  do
7:    temp = closure
8:    foreach ruleA → B ∈ F do
9:      if A ⊆ closure then
10:       closure = closure ∪ B
11:       rule = ∅
12:       foreach item ∈ A do
```

13: $rule = rule \cup rules[item]$
14: **end**
15: **foreach** $item \in B$ **do**
16: $rules[item] = rule$
17: **end**
18: **end**
19: **end**
20: **while** $temp \neq closure$
21: **if** $Z \subseteq closure$ **then**
22: **return** $rules[Z]$
23: **else**
24: **return** $\emptyset$
25: **end**

Fig. 3. Algoritmo de cierre con incertidumbre.

D. Umbral de certeza

Una vez que el grado de redundancia se calcula para una regla es necesario determinar si es suficiente declarar la regla redundante o por el contrario si no hubiese evidencia suficiente sobre la redundancia. La manera más simple de abordar este problema es establecer un umbral definido por el usuario, por encima de este, es suficiente para considerar la regla como redundante.

V. EXPERIMENTOS

Se realizaron algunos experimentos para validar el método propuesto. Los datasets usados para este propósito se presentan en la Tabla 2. Todos están disponibles en el Repositorio de Aprendizaje Automático de la Universidad de California Irvine (UCI por sus siglas en inglés) [2]. Los registros en los datasets se convirtieron a transacciones y aquellas que tenían valores faltantes se eliminaron. Los atributos numéricos se transformaron usando un método de discretización de igual frecuencia con cuatro *bindings*.

TABLA II
DATASETS DE PRUEBA

Dataset	*Filas*	*Atributos*	*Atributos numéricos*
Adult	30162	15	6
Breast cancer	699	10	0
Car	1728	7	0
Credit	690	16	6
Haberman	306	4	2
Nursery	12960	9	0
Zoo	101	18	0

La medida usada para evaluar la propuesta es la tasa de reducción (BR por sus siglas en inglés) se define como el porcentaje de reglas eliminadas del modelo por redundancia del conocimiento previo. Las reglas de asociación se minaron usando el algoritmo A priori, los umbrales de soporte y confianza se definieron empíricamente para obtener conjuntos de reglas con un tamaño entre 1000 y 4000. El conocimiento del usuario se simuló usando un conjunto aleatorio de reglas con un tamaño igual al 1% del total de reglas minadas. La incertidumbre de las reglas fue simulada aleatoriamente asignando un valor entre $[0.75; 1]$. Se crearon 30 conjuntos de conocimiento para cada dataset. Una regla es considerada redundante si su grado de redundancia está por encima de 0.75. Los resultados presentados en la Tabla 3 contienen la tasa media de reducción para cada dataset. La primera columna denota el dataset usado para generar las reglas. La segunda el número de reglas generadas. La tercera presenta la alternativa de eliminación de redundancia utilizada, señalando que NU se utiliza para representar la eliminación de redundancia sin incertidumbre, DS para representar la eliminación de redundancia con incertidumbre y el uso del modelo DS para combinar las evidencias y CF denota la eliminación con incertidumbre y el modelo CF para combinar la evidencia. La cuarta columna muestra la tasa de reducción, es el valor medio entre 30 ejecuciones usando diferentes conjuntos de conocimiento.

En la Tabla 3 se puede apreciar que la reducción usando el factor de certeza es ligeramente mejor excepto en el caso del dataset "Adult" donde sí lo sobrepasa significativamente.

TABLA III
TASA DE REDUCCIÓN MEDIA PARA 30 EJECUCIONES SOBRE LOS DATASETS DE PRUEBA

Dataset	*Reglas*	*Método*	*Tasa de reducción (BR)*
Adult	2692	UN	95.6%
Adult	2692	CF	83.1%
Adult	2692	DS	47.7%
Breast cancer	3555	UN	93.7%
Breast cancer	3555	CF	84.8%
Breast cancer	3555	DS	84.2%
Car	2126	UN	15.1%
Car	2126	CF	12.0%
Car	2126	DS	12.0%
Credit	3429	UN	58.7%
Credit	3429	CF	40.9%
Credit	3429	DS	40.7%
Haberman	1179	UN	45.9%
Haberman	1179	CF	34.9%
Haberman	1179	DS	34.8%
Nursery	2136	UN	17.3%
Nursery	2136	CF	13.3%
Nursery	2136	DS	13.3%
Zoo	1352	UN	76.7%
Zoo	1352	CF	64.7%
Zoo	1352	DS	64.4%

VI. CONCLUSIÓN

La idea fundamental en este trabajo está vinculada a la definición principal de minería de datos: análisis de una gran cantidad de datos para extraer patrones interesantes, antes desconocidos y la consideración de que una regla de asociación que corresponde al conocimiento previo es redundante. Este enfoque presentado en este artículo poda aquellas reglas y presenta un modelo más simple al usuario final. La contribución fundamental de este trabajo es la definición de redundancia de reglas de asociación respecto al conocimiento previo impreciso, y la definición de un mecanismo para eliminar este tipo de redundancia utilizando el modelo de la teoría de la evidencia de Dempster-Schaffer y el modelo del Factor de Certeza. Los resultados alcanzados muestran que al utilizar el factor de certeza para propagar la incertidumbre en el modelo se alcanzan mejores resultados que cuando se utiliza la Teoría de Dempster-Schaffer como mecanismo para propagar la incertidumbre. Las diferencias más significativas se alcanzaron en el dataset Adult

donde el rendimiento logrado por el factor de certeza superó en más de un 40% al modelo DS. En el resto de los datasets las diferencias fueron mínimas a lo sumo dos unidades porcentuales.

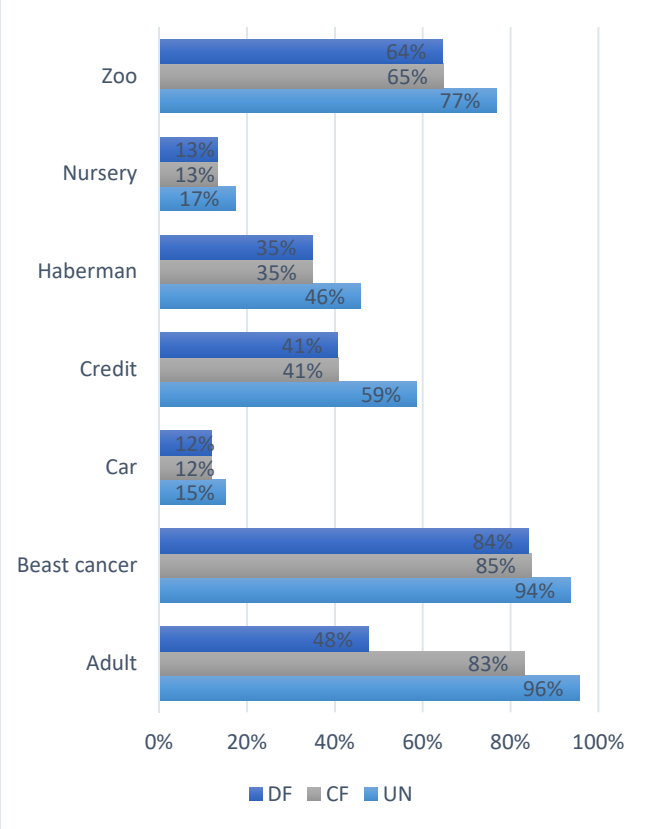

Fig. 4. Resultados de los experimentos.

REFERENCIAS

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. volume 22, pages 207–216. ACM, 1993.

[2] Asuncion and D J Newman. UCI Machine Learning Repository: Data Sets, 2015.

[3] José L. Balcázar. Redundancy, deduction schemes, and minimum-size bases for association rules. *arXiv preprint arXiv:1002.4286*, 2010.

[4] Yves Bastide, Nicolas Pasquier, Rafik Taouil, Gerd Stumme, and Lotfi Lakhal. Mining minimal non-redundant association rules using frequent closed itemsets. In *Computational LogicCL 2000*, pages 972–986. Springer, 2000.

[5] Niket Bhargava. Survey of Interestingness Measures for Association Rules Mining: Data Mining, Data Science for Business Perspective. *IRACST –International Journal of Computer Science and Information Technology & Security*, 6(2):74–80, 2016.

[6] Deborah Carvalho, Alex Freitas, and Nelson Ebecken. Evaluating the correlation between objective rule interestingness measures and real human interest. *Knowledge Discovery in Databases: PKDD 2005*, pages 453–461, 2005.

[7] A. P. Dempster and G. Shafer. Classic Works of the Dempster-Shafer Theory of Belief Functions, volume 219 of Studies in Fuzziness and Soft Computing. Springer, New York, 2008.

[8] Julio Diaz Vera, Carlos Molina Fernández, and María-Amparo Vila Miranda. Reducción de Redundancia en Reglas de Asociación. *Revista Cubana de Ciencias Informáticas*, 10(1):55–70, 2016.

[9] Michael Hahsler and Radoslaw Karpienko. Visualizing association rules in hierarchical groups. *Journal of Business Economics*, 87(3):317–335, Apr 2017.

[10] P. Krause and D. Clark. The Certainty Factor Model. Springer, Dordrecht, 1993.

[11] K. B. Laskey and T. S. Levitt. Artificial Intelligence: Uncertainty. In *International Encyclopedia of the Social and Behavioral Sciences*, pages 799 – 805. Elsevier, Oxford, U.K, 2001.

[12] D Maier. *The Theory of Relational Databases*. Computer Science Press, Rockville, 1983.

Julio Díaz Vera was born in Villa Clara, Cuba, on January 16, 1978. He received his M.S. degree in Computer Science in 2007 from the University of Informatics Sciences. He is an assistant professor in the Department of Software Engineering of the University of Informatics Sciences since 2005. His current main research interests are in the fields of Data warehousing, Data Mining and Soft Computing.

Guillermo Manuel Negrín Ortiz is a professor at the University of Informatics Sciences, Havana, Cuba. He received a B.S. degree in Informatics Sciences from the University of Informatics Sciences in 2014. His research interests include Association Rules Mining, Data Warehousing, Databases and Operating Systems.

Carlos Molina was born in Granada, Spain, on October 15, 1979. He received his M.S. degree in Computer Science in 2002 and his Ph.D. in Computer Science in 2005, both from the University of Granada. He is an assistant professor in the Department of Computer Science of the University of Jaen since 2004. His current main research interests are in the fields of Multidimensional Model, Data Mining and Soft Computing.

Amparo Vila is a full professor in the Department of Computer Science and Artificial Intelligence at the University of Granada, where she leads the Intelligent Databases and Information Systems research group. She holds a Ph.D. in Mathematics from the University of Granada. Her main research interests are in the fields of database design, data mining, and mathematical theory.