Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

Tipo de artículo: Artículo original
Temática: Inteligencia Artificial
Recibido: 18/01/2015 | Aceptado: 04/09/2015

Reducción de Redundancia en Reglas de Asociación

Redundancy Reduction in Association Rules

Julio Diaz Vera[1*], Carlos Molina Fernández[2], María- Amparo Vila Miranda [3]

[1] Universidad de la Ciencias Informáticas. La Habana Cuba. jcdiaz@uci.cu
[2] Universidad de Jaén. Jaén España. carlosmo@ujaen.es
[3] Universidad de Granada. Granada España. vila@decsai.ugr.es

[*] Autor para correspondencia: jcdiaz@uci.cu

Resumen

El minado de reglas de asociación es uno de los campos más estudiados y aplicados en minería de datos. Los modelos descubiertos usualmente contienen un número de reglas demasiado grande. Esto reduce la capacidad de los especialistas para utilizar los mismos con vista a la toma de decisiones. Esta deficiencia se acentúa cuando hay presente reglas redundantes dentro del modelo. En este trabajo se propone una definición de redundancia que tiene en cuenta el conocimiento previo de los usuarios con respecto al dominio. Se desarrolla un método, en la etapa de post-procesamiento, para reducir la redundancia de los modelos de reglas de asociación. La propuesta permite encontrar modelos más compactos que facilitan su utilización en el proceso de toma de decisiones. Los experimentos realizados han mostrado niveles de reducción cercanos al 90% del modelo. Las reglas consideradas como conocimiento previo no superan el 10% de las presentes en el modelo original. El método desarrollado facilita la utilización de las reglas de asociación en la toma de decisiones y por tanto aumenta la eficiencia de la minería de reglas de asociación.

Palabras clave: Minería de reglas de asociación, redundancia en reglas de asociación, post-procesamiento de reglas de asociación.

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 sets of rules; so the analysis capability, from a user point of view, are dismissing. It is difficult to use the found model in order to help the decision-making process. The previous handicap is accentuated in presence of redundant rules in the final set. In this work a new definition of redundancy in association rules is proposed, based in user's prior knowledge. A post-processing method to eliminate this kind of redundancy, using association rules known by user is developed. Our proposal allows to find more compact models of association rules

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

to facilitate its use in the decision-making process. The developed experiments have shown reduction levels that exceed 90% of all generated rules, using prior knowledge always below 10%. So our method improves the efficiency of association rules mining and the utilization of discovered association rules.

Keywords: *Association Rule Mining, Redundant Rules, Post-processing of association rules*

Introducción

Las reglas de asociación han sido uno de los modelos de minería de datos más estudiados a los largo del tiempo. Su meta principal es encontrar relaciones desconocidas entre elementos de una base de datos. Una regla de asociación se presenta como una implicación de la forma $X \rightarrow Y$ donde X y Y se denominan antecedente y consecuente de la regla respectivamente. X y Y son conjuntos de ítems que satisfacen $X \cap Y = \emptyset$. Una regla de asociación refleja cuanto influye la presencia del antecedente de la regla en la presencia del consecuente para un registro de la base de datos.

La importancia de una regla es evaluada, de manera general, a partir de dos medidas estadísticas el soporte y la confianza. El soporte de una regla $sop(X \rightarrow Y)$ representa la porción de la base de datos para la que $X \cup Y$ es verdadero. La confianza $conf(X \rightarrow Y)$ representa la porción de registros que contienen Y dentro de aquellos que contienen X.

El descubrimiento de reglas de asociación facilita el proceso de toma de decisiones. Pero la cantidad de reglas resultantes limita la capacidad de los especialistas del negocio a la hora de procesar e interpretar las reglas, para aprovechar el conocimiento que se deriva de las mismas. Una parte importante de las reglas que se presenta a los usuarios es irrelevante debido a que son triviales, demasiado generales, demasiado específicas o porque no son significativas en la toma de decisiones. Esta limitación se acentúa en modelos que contienen reglas redundantes.

La comunidad científica ha introducido varias definiciones de redundancia en reglas de asociación. Desde la idea general planteada en (Bastide et al. 2000) "*una regla de asociación es redundante si cubre la misma información, o información menos general, que la información que cubre otra regla de la misma utilidad y relevancia*" hasta proposiciones formales como la también presentada en (Bastide et al. 2000) como Reglas de Asociación Mínimas no Redundantes. Definida como se muestra a continuación; una regla de asociación $R: X \rightarrow Y$ es mínima y no redundante si y solo si no existe una regla de asociación $R_1: X_1 \rightarrow Y_1$ tal que:

1. $sop(R) = sop(R_1)$
2. $conf(R) = conf(R_1)$
3. $X_1 \subseteq X \wedge Y_1 \subseteq Y$

Uno de los caminos más transitados por los investigadores para enfrentar el problema de la redundancia en reglas de asociación es encontrar un conjunto de reglas no redundantes que sea capaz de representar todas las reglas válidas dentro del dominio. En este sentido se han desarrollado dos variantes fundamentales a) utilizar criterios subjetivos o medidas estadísticas con vista a podar las reglas en la etapa de post-procesamiento y b) incrustar los mecanismos para la poda de las reglas dentro de los algoritmos de extracción y generación de reglas.

Las soluciones reportadas en la literatura no son suficientes desde el punto de vista del usuario, que continúa enfrentado modelos que contienen demasiadas reglas para poder ser interpretadas. Este trabajo propone una solución al problema de la redundancia en reglas de asociación. A partir del conocimiento previo del usuario sobre el dominio en cuestión se desarrolla un procedimiento en dos pasos. Primero se detectan y eliminan las reglas redundantes con respecto al conocimiento previo. En segundo lugar se detectan los elementos redundantes en el antecedente y/o el consecuente de una regla simplificando la misma. Como parte del desarrollo de este trabajo se adapta la definición de redundancia en reglas de asociación y se propone una base genérica para desarrollar el proceso de reducción de redundancia en la etapa de post-procesamiento.

Metodología computacional

Redundancia en reglas de asociación

Se han desarrollado varias alternativas para tratar el tema de la redundancia en reglas de asociación. El uso conjuntos de conceptos frecuentes fue propuesto en (Hui 2013) como mecanismo para podar un conjunto de reglas restringiendo de manera inherente las reglas con respecto a los objetos. Una línea que ha recibido atención es la creación de métricas de interés que permitan decantar una parte de las reglas encontradas (Lerman and Guillaume 2013; Watanabe 2010; Djenouri et al. 2014; Kryszkiewicz 2015).

Las taxonomías de conceptos definidas sobre los ítems en la base de datos es quizás la única forma de conocimiento previo utilizada para reducir el tamaño de los modelos de reglas de asociación. En esta línea se destacan los trabajos (Baralis et al. 2012; Kaoungku, Kerdprasop, y Kerdprasop 2014; Dimitrijevic y Bosnjak 2014). Es necesario acotar que la representación de conocimiento utilizada está limitada a las relaciones de tipo es-un que aparecen en los ítems lo que difiere de la propuesta planteada en este trabajo que permite utilizar cualquier tipo de relación entre los ítems de la base de datos.

Todas estas variantes comparten la misma deficiencia, se pierde algún tipo de información y la cantidad de reglas presentada al usuario final continúa siendo prácticamente imposible de asimilar. No obstante lograron demostrar que un conjunto relativamente pequeño de reglas puede ser presentado al usuario en lugar de todas las reglas encontradas, sin socavar la capacidad de toma de decisiones.

El problema de la redundancia en reglas de asociación tiene su génesis en (Toivonen et al. 1995) donde se planteó que los conjuntos de reglas de asociación descubiertos son altamente redundantes debido a que varias reglas describen las mismas filas de la base de datos. Él propuso un método de reducción de redundancia basado en reglas de cobertura. Las cuales representan un subconjunto de las reglas descubiertas pero esta propuesta al ser independiente del dominio no es capaz de podar las reglas que son previamente conocidas por el usuario.

Algunas de las variantes de reducción de redundancia se orientan hacia la construcción de mecanismos de inferencias a partir de los cuales se puedan generar el resto de las reglas. De esta forma se presentan un conjunto reducido de las reglas al usuario. En (Zaki 2004) se proponen las Reglas de Asociación Cerradas que se basan en los itemset frecuentes cerrados, pero la reducción que se alcanza puede ser muy pobre en juegos de datos dispersos. Como en el resto de los casos estudiados no es capaz de detectar las reglas previamente conocidas por el usuario.

En (Pasquier et al. 2005) se propone la base Min-Max que utiliza conectores de Galois para extraer reglas de asociación no redundantes desde los itemset frecuentes cerrados, en lugar de hacerlo desde los itemsets frecuentes. En (Prakash, Govardhan, y Sarma 2013) se propone el uso de esta base para obtener reglas aproximadas. La Base Genérica Informativa fue presentada en (Gasmi et al. 2005). Esta base incluye la particularidad de contar con el soporte de todos los itemset frecuentes por lo que puede calcularse el soporte y la confianza de todas las reglas generadas. Pero la base aún incluye demasiadas reglas lo que la hace poco compacta. En (Sahoo, Das, and Goswami 2014) se refina este mecanismo proponiendo la base genérica informativa compacta.

La Base Mínima Genérica fue desarrollada en (Cherif et al. 2005) esta base es construida a partir del retículo aumentado del Iceberg de Galois. En ese trabajo se introduce el concepto de base parcialmente informativa. En una base parcialmente informativa el soporte y la confianza pueden determinarse exactamente para un grupo de reglas mientras que en otras solo se determina el intervalo en el que se encontraran. Otro enfoque basado en generadores mínimos es propuesto en (Chen, Wang, and Cao 2014).

(Balcázar 2010) propone una definición de redundancia en reglas de asociación basada en el cierre y la define de la siguiente manera: Sea β un conjunto de implicaciones. Una regla de asociación $X_0 \rightarrow Y_0$ es redundante con respecto a β y la regla $X_1 \rightarrow Y_1$ denotado $\beta, \{X_1 \rightarrow Y_1\} \vDash X_0 \rightarrow Y_0$ si cualquier conjunto de datos D en el que todas las reglas en β

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

se cumplen con confianza 1 provoca que $conf(X_0 \rightarrow Y_0) \geq conf(X_1 \rightarrow Y_1)$. En el artículo se proponen tres esquemas de inferencia reducción a la derecha (rR), aumento a la derecha (rA) y aumento a la izquierda (lA).

- (rR) si $X \rightarrow Y \; and \; Z \subset Y \; entonces \; X \rightarrow Z$
- (rA) si $X \rightarrow Y \; entonces \; X \rightarrow XY$
- (lA) si $X \rightarrow YZ \; entonces \; XY \rightarrow Z$

Las reglas redundantes confiables fueron introducidas en (Xu, Li, and Shaw 2011) donde se desarrolló la base confiable conformada por dos bases, ConfiableAproximada usada para las reglas con confianza menor que 1 y ConfiableExacta usada en las reglas con confianza 1. Se utilizan itemsets frecuentes cerrados para llevar a cabo el proceso de reducción. (Pasquier et al. 2005) propone generar reglas con mínimo antecedente y máximo consecuente. Esta base logra eliminar una gran cantidad de redundancia pero una parte de las reglas presentadas al usuario final pueden ser inútiles debido a que forman parte del conocimiento previo del usuario.

Todos los trabajos discutidos hasta el momento de alguna forma presentan al usuario final un grupo de reglas que no tienen significado en la toma de decisiones debido a que ya las conoce, dicho de otra forma en estos modelos persisten reglas que son redundantes con respecto al conocimiento previo.

Redundancia basada en el conocimiento previo

Sea S un conjunto de reglas de asociación y S_c un conjunto de reglas previamente conocidas que se asumen como de confianza 1, definidas sobre el mismo dominio que S. Una regla de asociación $R: X \rightarrow Y \in S$ es redundante con respecto a S_c si existe una regla $R': X' \rightarrow Y'$ que satisface alguna de las siguientes condiciones:

1. $X' \subseteq X \bigwedge Y' \subseteq Y$
2. $X' \subseteq X \wedge \exists R'': X'' \rightarrow Y'' \in S_c: Y \subseteq Y''$
3. $X' \subseteq X \bigwedge Y \subseteq Y'$
4. $X' \subseteq X \bigwedge Y' \subseteq X$
5. $X' \subseteq Y \bigwedge Y' \subseteq Y$

Tabla 1. Dataset Binario

Tid	Ítems	Itemsets	Soporte
1	A,C,D	D	1/5
2	B,C,E	AB, AE, ABC, ABE, ACE, ABCE	2/5
3	A,B,C,E	A, AC, BC, CE, BCE	3/5
4	B,E	B, C, E, BE	4/5
5	A,B,C,E	{∅}	1

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

Ejemplo 1: En la tabla 1 se representa una base de datos D conjuntamente con los itemsets frecuentes y el soporte de los mismos. Asumiendo que las reglas de asociación $B \rightarrow C$ y $C \rightarrow E$ son suministradas como conocimiento previo por parte del usuario tal que $S_c = \{B \rightarrow C,\ C \rightarrow E\}$. Si se define un umbral de soporte y confianza del 40% el conjunto de las reglas de asociación fuertes derivadas de D contiene 49 reglas dentro de las que se encuentran:

$R = \{ A \rightarrow BC, \text{AB} \rightarrow \text{C}, B \rightarrow CA, \text{BC} \rightarrow \text{A}, \text{AB} \rightarrow \text{E}\}$

La regla A → BC presente en R satisface la condición 5, de la definición de redundancia planteada en este trabajo, con respecto a la regla B → C en Sc. Ya que $(B \subseteq B \wedge C \subset BC)$ por ello la regla es redundante con respecto al conocimiento previo. Si se toma la regla AB → C en R se puede comprobar que cumple la condición 3 con respecto a la regla $B \rightarrow C$ de Sc. Debido a que $C \subseteq C \wedge B \subset BC$ y de esta forma la regla es redundante con respecto al conocimiento previo. Mientras que la regla $B \rightarrow CA$ es redundante debido a que cumple con la condición 1 con respecto a la regla $B \rightarrow C$ debido a que $(B \subset B \wedge C \subseteq CA$. La regla BC → A cumplimenta la condición 4 con respecto a la regla B → C que forma parte del conocimiento previo dado (B ⊆ B ∧ C ⊆ C). Por último la regla AB → E satisface la condición 2 con respecto a la regla B → C tomada como R' y la regla $C \rightarrow E$ tomada como R'' debido a que

(B ⊂ AB ∧ E ⊆ E).

Los axiomas de Armstrong son un conjunto de reglas de inferencias que permiten encontrar el conjunto mínimo de dependencias funcionales que satisface una base de datos. A partir de estos axiomas pueden derivarse el resto de las dependencias funcionales. Este sistema de inferencia permite encontrar subconjuntos reducidos de dependencias funcionales denominados “coberturas” que son equivalentes a las “bases” en minería de datos.

(Balcázar 2010) descartó la posibilidad de utilizar los axiomas de Armstrong como mecanismo de inferencia en reglas de asociación debido a que es imposible obtener los valores de soporte y confianza para las reglas derivadas:

- Reflexividad si $B \subset A\ entonces\ A \rightarrow B$ se cumple en las reglas de asociación debido a que la $conf(A \rightarrow B) = \frac{sop(A \cap B)}{sop(A)} = \frac{sop(A)}{sop(A)} = 1$
- Transitividad si $A \rightarrow B$ y $B \rightarrow C$ no se cumplen, con un soporte mayor que un umbral no es posible conocer el $sop(A \rightarrow C)$
- El axioma aumentativo tampoco es factible ya que si $A \rightarrow B$ no puede garantizarse que $AC \rightarrow B$ se satisfaga. Aumentar el antecedente de la regla puede generar una regla con una menor confianza incluso al punto de

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

llegar a cero, lo que ocurriría en el caso donde la mayoría de las ocurrencias del ítem X en una base de datos estuviera acompañado por el ítem Z pero solo aparece en compañía del ítem Y cuando no aparece Z.

En este trabajo se propone el uso de los axiomas de Armstrong, no como mecanismo de inferencia de nuevas reglas sino para evaluar si una regla presenta redundancia con respecto al conocimiento previo representado como un conjunto de reglas conocidas S_c.

La propuesta que se realiza en este trabajo pretende reducir los elementos redundantes en el antecedente y consecuente de una regla. A pesar de que la aplicación de los axiomas de Armstrong no garantiza que las reglas derivadas cumplan la restricción de soporte. Gracias a la propiedad de clausura descendente del soporte se puede asegurar que todas las reglas que se deriven de la reducción del antecedente o el consecuente de una regla van a satisfacer el umbral de soporte de las reglas.

Algoritmos para eliminar la redundancia con respecto al conocimiento previo

En esta sección se presentan dos algoritmos el primero, ver algoritmo 1, permite reescribir las reglas que contienen ítems redundantes y el segundo, ver algoritmo 2, permite encontrar las reglas cuya información es redundante y pueden ser podadas. En ambos se utiliza el algoritmo de cierre presentado en (Maier 1983) para computar X^+.

Algoritmo 1. Reescritura de reglas

Entrada: Conjunto de reglas previamente conocidas S_c
Regla a ser reescrita $R_i$en la forma $X \rightarrow Y$
Salida: Regla reducida R_i'

$F = R_c \cup \{R_i\}$
/Reducir antecedente/
For all elemento A in X **do**
  If$((X - \{A\})^+ \; over \; F = \; (X - \{A\})^+ \; over \; ((F - \{R_i\}) \cup \; (X - \{A\}) \rightarrow Y))$ **then**
    $R_i = \; X - \{A\} \rightarrow Y$
  End if
End for
If $(conf(R_i) \leq threshold)$ **then**
  $prune(R_i)$
End if
/Reducir consecuente/
For all elemento W in Y **do**
  If$(X^+ overF = \; X^+ over \; ((F - \{R_i\}) \cup \; (X \rightarrow Y - \{W\}))$ **then**
    $R_i = \; X \rightarrow Y - \{W\}$
  End if
End for
Return R_i

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

Para eliminar un elemento A del antecedente de una regla es necesario obtener ese elemento a partir del conocimiento previo. La primera parte del algoritmo 1 verifica esta propiedad para todos los elementos del antecedente de la regla analizada. Se computa el cierre del nuevo antecedente $X\backslash\{A\}$ sobre el conjunto de reglas conocidas unido a la regla analizada. Luego se compara el resultado con el cierre de $X\backslash\{A\}$ computado sobre un nuevo conjunto de reglas en las que se sustituye la regla analizada por la regla $X\backslash\{A\} \rightarrow Y$. Si los cierres coinciden el elemento A es eliminado del antecedente de la regla. En este caso debe verificarse que la regla resultante satisface el umbral de confianza definido por el usuario.

Una regla redundante con respecto al conocimiento previo puede ser podada si el consecuente de la misma puede derivarse a partir del conjunto de reglas conocidas. Esta característica puede verificarse computando el cierre del antecedente de la regla sobre el conjunto de reglas en el conocimiento previo. Si el consecuente de la regla es un subconjunto del cierre entonces la regla puede ser podada. El algoritmo 2 desarrolla el procedimiento necesario para realizar esta tarea.

Algoritmo 2. Poda de Reglas

Entrada: Conjunto de reglas conocidas S_c
Regla analizada R_i en la forma $X \rightarrow Y$
Salida: Valor boolenao: **true** si la regla es podada **false** en otro caso.

$F = R_c$
If $(Y \in (X)^+ \; over \; F - \{X \rightarrow Y\})$ **then**
Return true
Else
Return false
End if

La complejidad temporal de los algoritmos es una función $T(n)$ que constituye una cota superior para el número de operaciones máximas de un algoritmo para una entrada de tamaño n. El modelo RAM (máquina de acceso aleatorio por sus siglas en inglés) es el más utilizado y representa una computadora digital simple con acceso aleatorio a memoria. En aras de la simplicidad $T(n)$ es aproximada por una función más sencilla, denotada como $T(n) = O(f(n))$ si existen las constantes $c \geq 0$ y $n_1 \geq 0$ tal que $T(n) \leq cf(n)$ para toda $n \geq n_1$.

En el algoritmo 1 se considera como a al número de símbolos presentes en S_c y como p al número de reglas que forman parte del conjunto de reglas previamente conocidas S_c. El orden de complejidad para computar el cierre es $O(n)$ tal como se detalla en (Maier 1983). El tiempo de ejecución del primer **for** en el algoritmo 1 es $a * 2p$ debido a que el número de reglas en F y el número de reglas en F' es p y se computa dos veces el cierre con un costo de $O(p)$. Este

valor se puede aproximar como $O(ap)$ ya que la constante 2 puede ser ignorada. El tiempo de ejecución del segundo ciclo **for** toma el mismo valor $a * 2p$ debido a que en él se realizan las mismas operaciones. Para computar la complejidad del algoritmo 1 deben sumarse los valores de complejidad de los dos ciclos lo que genera $O(ap) + O(ap) = 2O(ap)$ pero como el valor de la constante puede descartarse el valor final sería $O(ap)$.

La complejidad temporal del algoritmo 2 es mucho más simple ya que solo realiza una llamada al algoritmo de cierre y por tanto su complejidad temporal es $O(p)$. La complejidad temporal de todo el proceso es la suma de la complejidad del algoritmo 1 y el algoritmo 2 $O(ap) + O(p)$ que puede ser aproximada al mayor de los valores por lo que finalmente sería $O(ap)$. Los valores de a y p son normalmente mucho menores que la cantidad de transacciones y los ítems en la base de datos.

Los algoritmos de extracción de reglas de asociación tienen un orden de complejidad superior (Singh, Agarwal, y Rana 2013) al del procedimiento de reducción presentado en este trabajo. Esta diferencia sustenta la decisión de utilizar el procedimiento de reducción como parte del post procesamiento. De forma tal que se ejecute una vez el algoritmo de extracción y a continuación se ejecute el proceso de reducción teniendo en cuenta el conocimiento previo de varios usuarios sin necesidad de lanzar una vez más el algoritmo de extracción de reglas.

Resultados y discusión

La validez de la propuesta es verificada mediante la aplicación del procedimiento de reducción en tres juegos de datos. El primero con información demográfica proveniente del censo de población y vivienda de los Estados Unidos de América (Blake y Merz 2007), el segundo con información referente a la inversión en bolsa (Núñez 2007) y por último un dataset con datos hipotéticos sobre especies de hongos (Blake y Merz 2007). Para cada caso se utiliza como conocimiento previo un conjunto de seis reglas. Es necesario acotar que el objetivo del experimento tiene carácter didáctico centrándose en demostrar la aplicabilidad del método de reducción. Es claro que los resultados dependerán en gran medida de la calidad del conocimiento previo. Las reglas utilizadas en los experimentos responden al sentido común como por ejemplo: $\{[\text{Relacion}].[\text{Relacion}].[\text{Husband}] \rightarrow [\text{EstadoCivil}].[\text{Estado Civil}].[\text{Married} - \text{civ} - \text{spouse}]\}$ que representa el hecho de que si la relación de una persona con el propietario de una vivienda es esposo entonces esa persona tiene como estado civil casado. Esta es una regla que se utiliza como conocimiento previo en el dataset Censo.

Para evaluar los resultados alcanzados se utilizan dos métricas:

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

1. Ratio de poda: $P_r = \frac{ReglasPodadas}{TotalReglas} * 100$ esta es la métrica fundamental y define en última instancia la efectividad de la propuesta.
2. Ratio de reducción: $R_R = \frac{ReglasSimplificadas}{TotalReglas} * 100$ una métrica auxiliar que denota la simplificación de reglas.

El tiempo de ejecución no es presentado porque es despreciable en comparación con el tiempo de ejecución de los algoritmos de extracción de reglas.

La tabla 2 muestra los resultados alcanzados. Cada fila corresponde a un experimento con los siguientes pasos:

- Encontrar el conjunto de reglas de asociación con soporte superior al especificado en la columna 2 y confianza superior a la de la columna 3 la cantidad de reglas encontradas se reflejan en la columna 4.
- Aplicar el procedimiento de reescritura de reglas descrito en el algoritmo 1. El número de reglas simplificadas es recogido en la columna 5.
- Aplicar el procedimiento de poda de reglas descrito en el algoritmo 2. El número de reglas podadas es presentado en la columna 6.
- La cantidad final de reglas se presentan en la columna 7 mientras que las columnas 8 y 9 almacenan el ratio de reducción y el de poda respectivamente.

Tabla 2. Resultados Experimentales

Dataset	**Soporte**	**Confianza**	**Reglas**	**Simplificadas**	**Podadas**	**Finales**	**RatioReducción**	**RatioPoda**
Censo	0.01	0.4	3408	1119	942	2466	45	27
Censo	0.03	0.4	835	303	242	593	51	28
Censo	0.05	0.4	458	190	158	300	70	32
Censo	0.07	0.4	229	94	79	150	62	34
Censo	0.09	0.4	163	60	51	112	53	31
Censo	0.11	0.4	114	34	23	91	37	20
Bolsa	0.2	0.4	11010	7229	5592	5418	65	50
Bolsa	0.3	0.4	3314	1591	2225	1089	48	67
Bolsa	0.4	0.4	1230	307	904	326	24	73
Bolsa	0.5	0.4	349	46	294	55	13	84
Bolsa	0.6	0.4	212	65	64	148	30	30
Hongos	0.3	0.5	78998	34655	29154	49844	74	36
Hongos	0.4	0.5	5767	2484	1225	4542	43	21
Hongos	0.5	0.5	1148	439	200	948	38	17
Hongos	0.6	0.5	266	74	88	178	27	33
Hongos	0.7	0.5	180	64	83	97	35	46

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

El ratio de poda varía de acuerdo a las variaciones del soporte utilizado en el proceso de extracción. En el caso del Censo y la Bolsa primero aumentan en la medida que los valores de soporte son mayores pero, cuando el soporte alcanza un umbral de 0.07 en el caso del Censo y de 0.5 en el de la bolsa, el comportamiento cambia y comienza a decrecer. Para el dataset Hongos el comportamiento es inverso el ratio decrece a medida que aumenta el soporte hasta que se alcanzan soportes mayores de 0.5 donde el ratio crece directamente proporcional al soporte.

El comportamiento antes referido refleja una relación entre el soporte y los patrones de conocimiento previo. Cuando el valor de soporte aumenta un grupo de reglas deja de satisfacer la condición de umbral, definida por el usuario para la extracción, y no son incluidas en el modelo. Si las reglas descartadas no tienen gran impacto sobre el total de reglas incluidas en el modelo, que pueden ser derivadas del conocimiento previo, el Ratio de Poda crecerá en la medida que crece el soporte. Cuando se alcance un punto para el cual el aumento de soporte empiece a eliminar las reglas derivadas del conocimiento del modelo final en este caso el Ratio de Poda comenzará a disminuir en la medida que el soporte es incrementado.

En la figura 1 se puede apreciar este comportamiento cuando se aplica el procedimiento de reducción con distintas cantidades de reglas en el conocimiento previo, representando el valor medio de reducción en cada grupo.

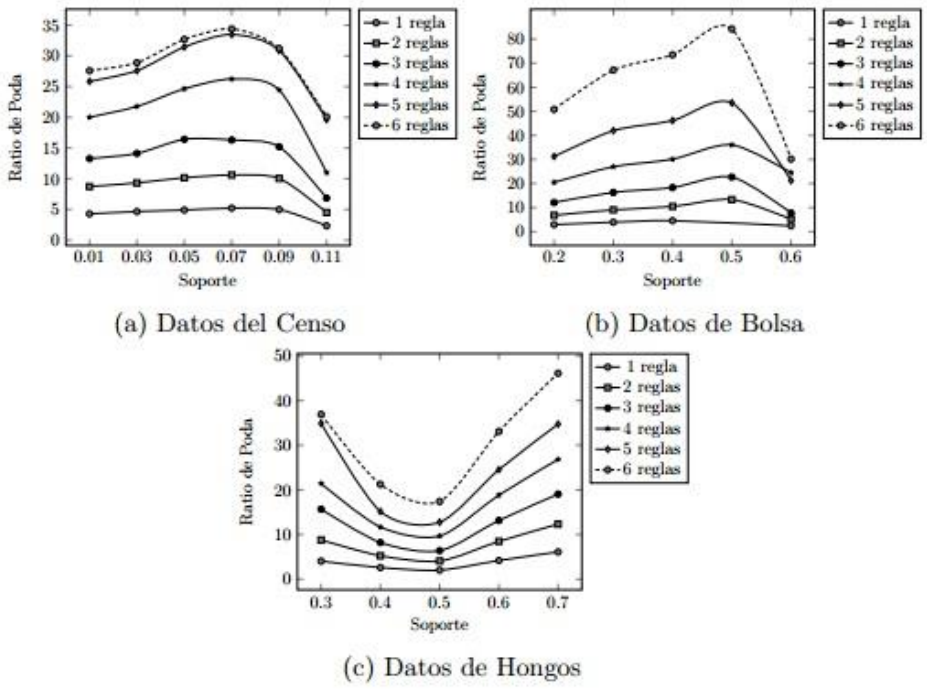

Figura 1. Reglas podadas

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

Comparación con enfoques tradicionales

Los enfoques anteriores a este trabajo abordan el problema de la redundancia en reglas de asociación centrándose en las relaciones estructurales de las reglas y diseñando mecanismos de reducción basados en reglas de inferencia o itemsets maximales. En este trabajo se potencia la utilización del conocimiento previo que disponen los expertos del negocio para eliminar aquellas reglas que no son capaces de brindar nuevo conocimiento. En esencia ambas variantes no son comparables pero se ha desarrollado un experimento con vista a comprobar si es posible alcanzar ratios de reducción similares a los reportados en la literatura a partir del procedimiento descrito en esta investigación. Debido a que no se han encontrado en la literatura consultad trabajos que enfoquen la reducción de redundancia a partir de conocimiento previo.

En la figura 2 se presenta el ratio de reducción alcanzado por algunas de las técnicas de reducción de redundancia más significativas, aplicadas sobre el dataset Hongos con un soporte de 0.3. Ha sido escogido este dataset debido a que es el único para el que están disponibles los resultados experimentales de los autores y es suficiente para establecer la comparación. Los valores para los ratios de reducción son los dados en los artículos originales de los autores: MinMax (Pasquier et al. 2005), Reliable (Xu, Li, and Shaw 2011), GB (Bastide et al. 2000), CHARM (Zaki 2004), CRS (Liu, Liu, and Zhang 2011) y Meta-Rules (Berrado and Runger 2007).

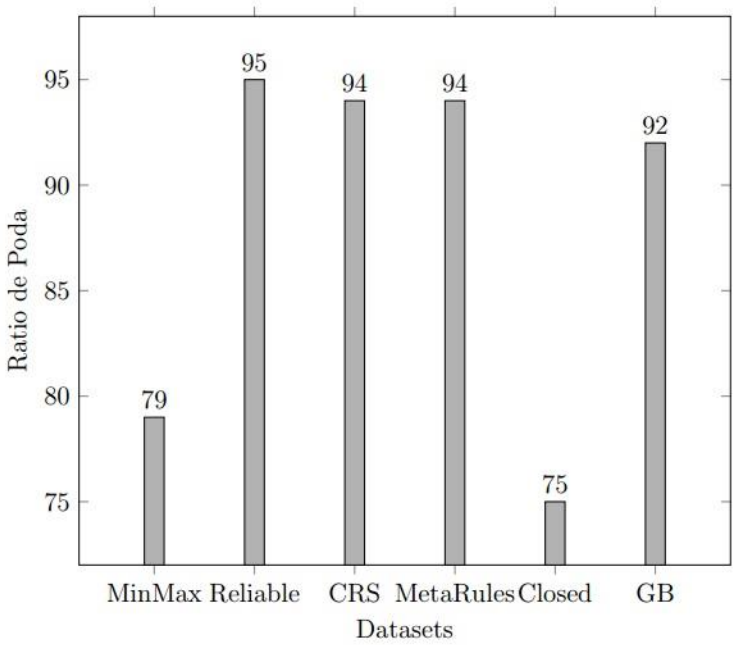

Figura 2. Ratio de Poda para las variantes previas

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 55-70

Reliable tiene el mejor Ratio de Poda de todas las variantes encontradas en la literatura. Esa es la razón por la que se decide comparar con ella los resultados alcanzados en este trabajo. La comparación se realiza para distintos valores de soporte en la extracción de reglas. En la tabla 3 se muestran los resultados. Se denota como $KBR_{Xreglas}$ a la aplicación del procedimiento de reducción descrito en este trabajo con X reglas en el conocimiento previo.

Tabla 3. Ratio de Poda

Soporte	**Reliable**	$\boldsymbol{KBR_{6reglas}}$	$\boldsymbol{KBR_{9reglas}}$	$\boldsymbol{KBR_{12reglas}}$	$\boldsymbol{KBR_{15reglas}}$
0.3	95	36	76	80	96
0.4	90	21	37	47	84
0.5	89	17	30	44	93
0.6	74	33	40	62	97
0.7	78	46	46	75	100
Promedio	85	32.5	45.8	61.5	94

A medida que la cantidad de reglas que forman parte del conocimiento previo aumentan el Ratio de Poda que se alcanza en la variante presentada en este trabajo se acerca a los valores de Reliable llegando a superarla en todos los casos a excepto cuando el soporte toma valor de 0.4, ver figura 3, cuando se utilizan 15 reglas como parte del conocimiento previo. Es necesario señalar que en este caso cuando se utilizan 15 reglas en el conocimiento previo estas solo representan un 0.018% del total de reglas para el caso de un soporte de 0.3 y un 7.9% para el caso del soporte 0.7.

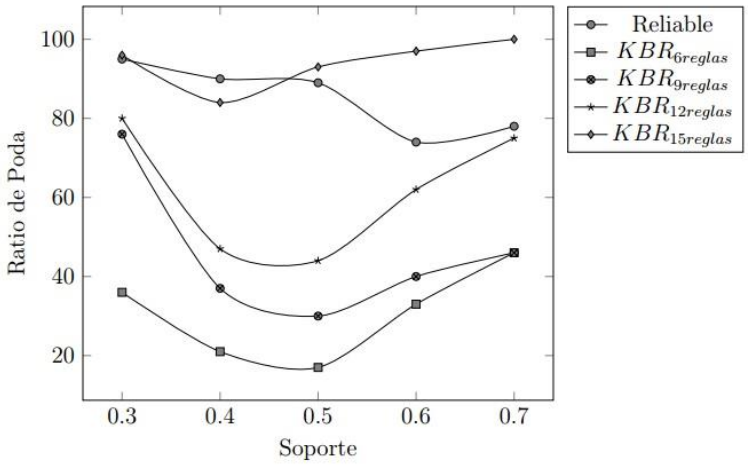

Figura 3. Ratio de Poda de Reliable vs KBR

KBR es capaz de mejorar el Ratio de Poda de las variantes previas utilizando una porción muy pequeña de reglas como conocimiento previo. Es claro que existe una relación muy estrecha entre el Ratio de Poda y la repercusión que tiene el conocimiento previo sobre el conjunto total de reglas. El Ratio de Poda aumentara en la misma medida en que el

conocimiento previo sea capaz de describir el dominio bajo estudio de esta forma se alcanzará mayor grado de reducción en la medida en que los usuarios conozcan mejor el dominio. El procedimiento presentado tiene además la capacidad de determinar cuándo un modelo no puede ser mejorado como ocurre en la aplicación del proceso de reducción para un soporte de 0.7 y con 15 reglas en el conocimiento previo donde se logra podar el 100% de las reglas lo que señala que el conocimiento previo cuenta con toda la información relevante en el dominio.

Conclusiones

La idea fundamental de este trabajo esta enlazada con la definición de minería de datos: análisis de grandes volúmenes de datos para extraer patrones interesantes, previamente desconocidos. La propuesta presentada permite podar aquellos patrones que ya forman parte del conocimiento de los expertos del negocio.

La contribución fundamental es la definición de redundancia con respecto al conocimiento previo en reglas de asociación y el desarrollo de un mecanismo que permite eliminar este tipo de redundancia de los modelos de reglas de asociación. Tarea que se realiza en dos procedimientos el primero detecta y elimina elementos redundantes en el antecedente o el consecuente de una regla y el segundo determina si una regla es redundante en su totalidad.

Los resultados de este estudio confirman que es posible utilizar el conocimiento de los expertos para reducir el volumen de los modelos de reglas de asociación. Modelos con menos reglas pueden ser interpretados de manera más fácil por parte de los especialistas lo que contribuye a su utilización en la toma de decisiones. Los resultados experimentales arrojaron que a partir de un conocimiento previo que contiene menos del 10% de las reglas en el modelo original se pueden alcanzar Ratios de Poda superiores al 90%.

Referencias

BALCÁZAR, JOSÉ L. Redundancy, Deduction Schemes, and Minimum-Size Bases for Association Rules. Logical Methods in Computer Science 2010 6(2-3) pp.1–33.

BARALIS, ELENA, LUCA CAGLIERO, TANIA CERQUITELLI, PAOLO GARZA Generalized Association Rule Mining with Constraints. Information Sciences 2012 pp. 194: 68–84.

BASTIDE, Y, PASQUIER, N, TAOUIL, R, STUMME, G, LAKHAL Mining Minimal Non-Redundant Association Rules Using Frequent Closed Itemsets. *In* Computational Logic—CL 2000 pp. 972–986

BERRADO, ABDELAZIZ, GEORGE C. RUNGER Using Metarules to Organize and Group Discovered Association Rules. Data Mining and Knowledge Discovery 2007 14(3) pp. 409–431.

BLAKE, C.L, C.J MERZ, UCI Repository of Machine Learning Databases. University of California, Irvine, CA 2007. http://www.ics.uci.edu/m~learn/MLRepository.html.

CHEN, XIAO-MEI, CHANG-YING WANG, HAN CAO Association Rules Mining Based on Minimal Generator of Frequent Closed Itemset. *En* Ecosystem Assessment and Fuzzy Systems Management 2014 Pp. 275–282.

CHERIF, CHIRAZ LATIRI, W. BELLEGUA, S. BEN YAHIA, GHADA GUESMI VIE_MGB: A Visual Interactive Exploration of Minimal Generic Basis of Association Rules. *In* Proc. of the Intern. Conf. on Concept Lattices and Application (CLA 2005) pp. 179–196.

DIMITRIJEVIC, MAJA, ZITA BOSNJAK Pruning Statistically Insignificant Association Rules in the Presence of High-Confidence Rules in Web Usage Data. Knowledge-Based and Intelligent Information & Engineering Systems 18th Annual Conference, KES-2014 Gdynia, Poland, pp. 271–280.

DJENOURI, Y., Y. GHERAIBIA, M. MEHDI, A. BENDJOUDI N. NOUALI-TABOUDJEMAT An Efficient Measure for Evaluating Association Rules. *In* 6th International Conference of Soft Computing and Pattern Recognition 2014 pp. 406–410.

GASMI, GH, S. BEN YAHIA, E. MEPHU NGUIFO, YAHYA SLIMANI A New Informative Generic Base of Association Rules. *En* Advances in Knowledge Discovery and Data Mining pp. 81–90.

Hui, Wang, Ming Non-Redundant Associations From the Frequent Concept Sets on FP-Tree. TELKOMNIKA Indonesian Journal of Electrical Engineering 2013 11(7) pp 3604–3610.

KAOUNGKU, N, KERDPRASOP, K, KERDPRASOP, N A Technique for Association Rule Mining in Multiple Datasets. Electrical Engineering and Information Technology 2014 pp. 63: 241.

KRYSZKIEWICZ, MARZENA Dependence Factor for Association Rules. E*n* Intelligent Information and Database Systems. pp. 135–145. Lecture Notes in Computer Science.

LERMAN, ISRAËL CÉSAR, AND SYLVIE GUILLAUME Comparing Two Discriminant Probabilistic Interestingness Measures for Association Rules. *En* Advances in Knowledge Discovery and Management2013 pp. 59–83.

LIU, HUAWEN, LEI LIU, AND HUIJIE ZHANG A Fast Pruning Redundant Rule Method Using Galois Connection. Applied Soft Computing 2011 11(1): 130–137.

MAIER, DAVID 1983 The Theory of Relational Databases, vol.11. Computer science press Rockville.

NÚÑEZ, J.F Empleo de Fuzzy OLAP Para Obtener Reglas Que Caractericen Estrategias de Inversión. UGR.

PASQUIER, N, RAFIK T, BASTIDE, Y, STUMME, G, LAKHAL, L Generating a Condensed Representation for Association Rules. Journal of Intelligent Information Systems 2005 24(1): 29–60.

PRAKASH, R. VIJAYA, A. GOVARDHAN, SSVN SARMA Discovering Non-Redundant Association Rules Using MinMax Approximation Rules. (IJCSE) 2013 3(6).

SAHOO, J, ASHOK KUMAR DAS, A. GOSWAMI An Effective Association Rule Mining Scheme Using a New Generic Basis. Knowledge and Information Systems 2014 43(1): 127–156.

SINGH, ARCHANA, JYOTI AGARWAL, AJAY RANA Performance Measure of Similis and FP-Growth Algorithm. International Journal of Computer Applications 2013 62(6) pp 25–31.

Toivonen, H, Klemettinen, M, Ronkainen, Kimmo Hätönen Pruning and Grouping Discovered Association Rules. 1995.

WATANABE, TOSHIHIKO An Improvement of Fuzzy Association Rules Mining Algorithm Based on Redundacy of Rules. *En* Aware Computing (ISAC), 2010 2nd International Symposium on Pp. 68–73. XU, YUE, YUEFENG LI, GAVIN SHAW Reliable Representations for Association Rules. Data & Knowledge Engineering 2011 70(6) pp. 555–575.

ZAKI, MOHAMMED J. Mining Non-Redundant Association Rules. Data Mining and Knowledge Discovery 2004 9(3) pp. 223–248.