Revista Venezolana de Computación
ISSN: 2244-7040
http://www.svc.net.ve/revecom
Vol. 2, No. 1, pp. 48-59, Diciembre 2015
Selección de los Mejores Artículos de CoNCISa 2015

Un Mecanismo de Respuestas a Consultas en Presencia de Nulos

Josué Ramírez[1], Leonid Tineo[1]
ramirezjosue@usb.ve, leonid@usb.ve

[1]Departamento de Computación y Tecnología de la Información, Universidad Simón Bolívar, Caracas, Venezuela

Resumen: En cualquier descripción del mundo real se pueden distinguir modelos de tratamiento para la imprecisión a nivel de los atributos: valores probabilísticos, valores difusos, distancias, solidez y completitud, y valores nulos. Varios trabajos previos centran sus esfuerzos en dar un tratamiento adecuado a los valores nulos. Algunos de ellos los ignoran bajo el argumento de no tener suficiente significancia estadística. Otros usan técnicas de reescritura tratando de evitar, en lo posible, condiciones que involucran valores nulos. Una estrategia muy conocida es la llamada "imputación" que consiste en cambiar los valores nulos en la base de datos por valores sustitutos elegidos con algún tipo de criterio establecido. Usando técnicas de inferencia es posible hacer estimación de valores en lugar de los nulos. Cualquiera de estos mecanismos tiene sus desventajas como pueden ser muy poca aplicabilidad práctica, inviabilidad de aplicación o introducción de ruido. Específicamente el presente trabajo plantea la aplicación de un método en bases de datos relacionales, en donde no se realiza la imputación, sino una estimación en forma dinámica en el momento de ejecutar una consulta que involucre un atributo que tenga valores nulos. De esta manera se evitan las desventajas antes descritas. Por otro lado, para la estimación del valor que será usado en la consulta se plantea el uso de reglas de asociación difusa, técnica ésta que permite obtener un conjunto de reglas para estimar los valores nulos en función de los demás valores presentes en los registros y cuya aplicación se extiende a atributos con valores cuantitativos y no sólo categóricos.

Palabras Clave: Valores Nulos; Reglas Asociación Difusas; Procedimientos Almacenados; Bases de Datos Incompletas; Consultas Relacionales.

Abstract: In any description of the real world can be distinguished treatment models imprecision level attributes: probabilistic values, fuzzy values, distances, soundness and completeness, and nulls. Several previous studies focus its efforts on providing adequate treatment for null values. Some of them ignore them on the grounds of not having enough statistical significance. Others use rewriting techniques trying to avoid, if possible, conditions involving null values. A well-known strategy is called "imputation" that is to change nulls in the database by substitutes chosen values with some sort of criteria established values. Using inference techniques it is possible to estimate values instead of zero. Any of these mechanisms has its disadvantages such as very little practical applicability, infeasibility of application or introduction of noise. Specifically, this paper presents the application of a method in relational databases, where the imputation is not made, but an estimate dynamically when running a query involving an attribute that has nulls. In this way the disadvantages described above are avoided. Moreover, for estimating the value to be used in the query it raises fuzzy association rules, this technique is for obtaining a set of rules for estimating zero depending on other values in the registers and whose application extends to attributes with quantitative values and not just categorical.

Keywords: Null Values; Fuzzy Association Rules; Store Procedures; Incomplete Data Bases; Relational Queries.

I. INTRODUCCIÓN

Las bases de datos incompletas constituyen un problema común en muchos contextos. Especialmente en las bases de datos relacionales se modela la incompletitud a través de la pseudodescripción denominada nulo [1]. Este valor tiene asociada diversas interpretaciones y se han realizado diversos esfuerzos o trabajos para resolver el tratamiento del mismo.

Algunos trabajos proponen extensiones teóricas al modelo relacional con la intención de aceptar la existencia del nulo como estrategia para el manejo de la imprecisión [2][3][4].

Sin embargo, estos trabajos no han sido de mucha aplicación práctica.

También se ha propuesto ignorar los nulos, lo cual resulta de mayor sencillez, pero tampoco es de mucha utilidad práctica, pues su aplicabilidad está muy limitada. Para ignorar la presencia de valores nulos, se requiere que la proporción de los mismos no sea significativa estadísticamente, condición que no siempre se cumple [5].

Otros desarrollos consideran las estrategias para estimar los nulos y existen varias propuestas en la literatura. Algunas suponen el cumplimiento de ciertas restricciones que deben satisfacer los datos para garantizar buenos resultados [6][7] [8]. Otras técnicas que han resultado efectivas resultan complejas de implementar o se usan en ámbitos muy específicos o proveen soluciones difíciles de implementar o auditar [9][10][11][12][13][14].

En cuanto a la sustitución del valor nulo por una estimación, lo cual se conoce como imputación, no es fácil seleccionar la técnica correcta porque esto depende de múltiples factores y además no existen reglas específicas que guíen la escogencia o bien una vez que ésta se realiza es importante también aplicar correctamente la técnica para evitar resultados erróneos [5][15][16].

En este artículo se propone un mecanismo automatizado que permite dar respuesta a consultas en base de datos relacionales en presencia de nulos, mediante estimación dinámica, subsanando las dificultades mencionadas de las estrategias y técnicas antes descritas.

En la sección II se describen las áreas de investigación relacionadas con nuestra propuesta. En la sección III se detalla la solución propuesta. En la sección IV se explican los experimentos y resultados más relevantes del sistema desarrollado. Por último, se presentan las conclusiones del artículo y las líneas de trabajo futuras.

II. BASES TEORÍCAS

A continuación se describen las áreas de investigación que sirvieron de base para la propuesta planteada en este artículo.

A. Reglas de Asociación

El modelo de reglas de asociación es muy conocido en su versión clásica para la obtención de patrones nuevos y significativos en minería de datos. Según [17], el modelo clásico de reglas de asociación concibe que el *ítem* es el objeto básico de interés. Un conjunto de *ítems* se conoce como *itemset*. Por otro lado una transacción es un hecho compuesto que agrupa un conjunto de ítems. Intuitivamente, las reglas son asociaciones que vinculan la presencia de *ítems* dentro de transacciones.

Más formalmente se puede definir una regla de asociación de la siguiente manera: Sea *I* un conjunto de *ítems* (*itemset*) y *T* un conjunto de transacciones con *ítems* en *I*, ambos conjuntos se asumen finitos. Una regla de asociación es $A \cap C = \varnothing$ una expresión de la forma $A \Rightarrow C$, donde $A, C \subseteq I$, $A, C \neq \varnothing$ y Esta regla significa que cada transacción de *T* que contiene *A* también contiene *C*.

Se consideran dos medidas de interés para evaluar las reglas de asociación: el soporte y la confianza, las cuales se basan en el concepto de soporte de un *itemset*, que se interpreta como la proporción de ocurrencia conjunta de uno o más ítems en una transacción. Las técnicas de minería procuran descubrir (minar) reglas fuertes, esto es reglas con soporte y confianza mayores que los umbrales definidos por el usuario. Sin embargo, a fin de disminuir el número de reglas minadas y que éstas sean más interesantes en [17] se propuso usar el factor de certeza. El factor de certeza es la medida de la variación de la probabilidad de que C esté en una transacción cuando se consideran sólo las transacciones donde se encuentra A.

Los valores del factor de certeza están entre [-1,1]. Su valor es positivo si la asociación entre A y C es positiva, es 0 cuando son independientes, y es negativo si están asociados negativamente. Específicamente, un factor de certeza positivo mide el grado de incremento en la probabilidad de que C se encuentre en una transacción, dado que A se encuentra. Una interpretación similar puede realizarse para valores negativos del factor de certeza [18].

El uso de las reglas de asociación se ha extendido al realizarse adaptaciones al algoritmo básico propuesto por Agrawal para poder buscar reglas en bases de datos incompletas [19]. Las reglas de asociación se han usado en las bases de datos relacionales, en donde incluso se han propuesto varias arquitecturas que incorporan la minería de reglas de asociación. Además, entre otras aplicaciones se han usado para la estimación de valores nulos [4][20].

B. Reglas de Asociación Difusa

En un esfuerzo para obtener patrones más significativos y subsanar deficiencias ante la incertidumbre que suele presentarse en los datos, el modelo de reglas de asociación se extendió mediante aplicación de teoría de conjuntos difusos, surgiendo las reglas de asociación difusa [18].

Una transacción difusa en el modelo de reglas de asociación difusa [18] es definido como conjunto difuso. Esto es, un conjunto en que la membresía de sus elementos es gradual en el intervalo real [0,1]. Los conjuntos difusos permiten dar una interpretación numérica de términos vagos del lenguaje natural denominados etiquetas lingüísticas [21].

Para todo $i \in I$, $\tilde{\tau}(i)$ denota el grado de membresía de i en una transacción difusa $\tilde{\tau}$. Análogamente, para $A \subseteq I$,

$$\tilde{\tau}(A) = \min_{i \in A} \tilde{\tau}(i) \qquad (1)$$

Un conjunto de N transacciones difusas con que agrupan a M ítems puede representarse como una matriz de N×M donde cada entrada $\alpha_{lk} = \tilde{\tau}_l(i_k)$

En [18], las medidas de interés, como el soporte y la confianza, se extienden al caso difuso usando un enfoque semántico basado en la evaluación de sentencias cuantificadas.

Las reglas de asociación difusa, entre otras aplicaciones, se han utilizado para la estimación de nulos [22], mas no han sido integradas a un mecanismo de consultas a bases de datos relacionales.

C. *Modelo para la Extracción de Reglas Difusas Mediante Niveles de Restricción*

En [23] se desarrolla un modelo formal para representar y evaluar reglas de asociación, tanto precisas como difusas, generalizando de una manera natural las medidas de interés del caso preciso al difuso. Para la extracción de las reglas de asociación se evalúa la frecuencia de las cuatro posibles combinaciones de dos *itemsets*: antecedente y consecuente.

Otro concepto útil en este modelo es el α-corte de un conjunto difuso. Esto se refiere al conjunto preciso de los elementos con membresía mayor o igual a un nivel α. Una propiedad difusa en un universo U puede ser representada por un conjunto de realizaciones precisas y que en el caso particular de un conjunto difuso se referirían a los α-cortes, los cuales serán llamados RL (niveles de restricción) [24].

Definición 1. [24] Un RL-set Λ es un conjunto finito de niveles de restricción $\Lambda = \{\alpha_1, ..., \alpha_m\}$ verificando $1 = \alpha_1 > \alpha_2 > ... > \alpha_m > \alpha_{m+1} = 0$ para $m \geq 1$.

En términos generales, el RL-set de una propiedad atómica representada por un conjunto difuso A sería como en la definición siguiente:

Definición 2. [24] Sea A un conjunto difuso definido sobre el referencial X. Entonces el RL-set para A viene dado por:

$$\Lambda_A = \{A(x) \mid x \in X\} \cup \{1\} \tag{2}$$

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

Definición 3. [24] Sean P,Q dos propiedades difusas con RL-representaciones (Λ_P, ρ_P) , (Λ_Q, ρ_Q) . Entonces, $P \wedge Q$, $P \vee Q$, $\neg P$ son propiedades difusas representadas por $(\Lambda_{P \wedge Q}, \rho_{P \wedge Q})$, $(\Lambda_{P \vee Q}, \rho_{P \vee Q})$ y $(\Lambda_{\neg P}, \rho_{\neg P})$ respectivamente, donde

$$\Lambda_{P \wedge Q} = \Lambda_P \cup \Lambda_Q \tag{3}$$

$$\Lambda_{\neg P} = \Lambda_P \tag{4}$$

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

$$\begin{aligned} \rho_{P \wedge Q}(\alpha) &= \rho_P(\alpha) \cap \rho_Q(\alpha) \\ &= \{x \in X \mid P(x) \geq \alpha \wedge Q(x) \geq \alpha\} \end{aligned} \tag{5}$$

$$\rho_{\neg P}(\alpha) = \overline{\rho_P(\alpha)} = \{x \in X \mid \neg(P(x) \geq \alpha)\} \tag{6}$$

Donde $\overline{Y}$ es el complemento usual de un conjunto preciso Y.

Ahora en función de las definiciones anteriores se procede a ilustrar que forma tendría una tabla de contingencia, que en el modelo mencionado se denomina tabla-4ft. Sean los conjuntos difusos $\tilde{\Gamma}_A$ y $\tilde{\Gamma}_B$ definidos en un conjunto de transacciones difusas $\tilde{D}$ como $\tilde{\Gamma}_A(\tilde{\tau}) = \tilde{\tau}(A)$ y $\tilde{\Gamma}_B(\tilde{\tau}) = \tilde{\tau}(B)$ respectivamente. Se puede denotar sus *RL*-representaciones como $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}})$ y $(\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$.

Ahora se consideran los conjuntos de ítems A, B y sus respectivas combinaciones $A \wedge B$, $A \wedge \neg B$, $\neg A \wedge B$ y $\neg A \wedge \neg B$, las cuales forman una partición de la base de datos $\tilde{D}$. También, se pueden considerar los conjuntos difusos asociados definidos en $\tilde{D}$ con sus respectivas *RL*-representaciones: $(\Lambda_{\tilde{A} \wedge \tilde{B}}, \rho_{\tilde{A} \wedge \tilde{B}})$, $(\Lambda_{\tilde{A} \wedge \neg \tilde{B}}, \rho_{\tilde{A} \wedge \neg \tilde{B}})$, $(\Lambda_{\neg \tilde{A} \wedge \tilde{B}}, \rho_{\neg \tilde{A} \wedge \tilde{B}})$ y $(\Lambda_{\neg \tilde{A} \wedge \neg \tilde{B}}, \rho_{\neg \tilde{A} \wedge \neg \tilde{B}})$.

Para cada $\alpha \in \Lambda_X$, $\rho_Y(\alpha)$ es un conjunto preciso y se puede calcular su cardinalidad de la manera usual: $|\rho_Y(\alpha)|$. De forma análoga, para cada $\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}$ se puede definir $a_i = |\rho_{\tilde{A} \wedge \tilde{B}}(\alpha_i)|$, $b_i = |\rho_{\tilde{A} \wedge \neg \tilde{B}}(\alpha_i)|$, $c_i = |\rho_{\neg \tilde{A} \wedge \tilde{B}}(\alpha_i)|$ y $d_i = |\rho_{\neg \tilde{A} \wedge \neg \tilde{B}}(\alpha_i)|$. Cada una de estas cardinalidades formarían parte de la tabla-4ft asociada a α_i (ver Tabla I), denotada como $M_{\alpha_i} = 4\text{-}ft(\tilde{\Gamma}_A, \tilde{\Gamma}_B, \tilde{D}, \alpha_i)$:

Tabla I: Tabla-4ft Asociada a α_i

M_{α_i}	$\tilde{\Gamma}_B$	$\tilde{\Gamma}_{\neg B}$
$\tilde{\Gamma}_A$	a_i	b_i
$\tilde{\Gamma}_{\neg A}$	c_i	d_i

La generalización de las medidas de interés es como sigue:

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

$$sop(A) = \sum_{\alpha_i \in \Lambda_{\tilde{A}}} (\alpha_i - \alpha_{i+1}) \left(\frac{|\rho_{\tilde{A}}(\alpha_i)|}{|\tilde{D}|} \right) \tag{7}$$

Al considerar la parte derecha de la fórmula (7) y usar la tabla-4ft asociada a los *itemsets A* y *B* para el cálculo del nivel de restricción α_i se obtiene lo siguiente:

$$\frac{|\rho_{\tilde{A}}(\alpha_i)|}{|\tilde{D}|} = \frac{a_i + b_i}{a_i + b_i + c_i + d_i} \tag{8}$$

La fórmula (8) da el soporte de un *itemset* cuando la base de datos es precisa y $a_i + b_i + c_i + d = n$ es constante para cualquier restricción y representa el número de transacciones difusas en $\tilde{D}$. A continuación se muestran el soporte y la confianza de una regla difusa. Sean $A, B \subseteq I$ dos *itemsets* disjuntos y $(\Lambda_{\tilde{A}}, \rho_{\tilde{A}})$, $(\Lambda_{\tilde{B}}, \rho_{\tilde{B}})$ las *RL*-representaciones asociadas a los conjuntos difusos $\widetilde{\Gamma}_A$ y $\widetilde{\Gamma}_B$ en $\tilde{D}$. Entonces, el soporte de la regla difusa $A \rightarrow B$ en $\tilde{D}$ se define como:

$$Sop(A \rightarrow B) = sop(A \wedge B) \tag{9}$$

$$sop(A \wedge B) = \sum_{\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1}) \left(\frac{|\rho_{\tilde{A} \wedge \tilde{B}}(\alpha_i)|}{|\tilde{D}|} \right) \tag{10}$$

Y la confianza de la regla difusa $A \rightarrow B$ en $\tilde{D}$ se define como:

$$Conf(A \rightarrow B) = \sum_{\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1})(\Rightarrow_I (a_i, b_i)) \tag{11}$$

$$\Rightarrow_I (a_i, b_i) = \frac{|\rho_{\tilde{A} \wedge \tilde{B}}(\alpha_i)|}{|\rho_{\tilde{A}}(\alpha_i)|} \tag{12}$$

Finalmente, se indica el factor de certeza. Sean $A, B \subset I$ dos *itemsets* disjuntos en *D* y sea $4ft\ A, B, D = \langle a, b, c, d \rangle$ su tabla-4ft asociada. El cuantificador $\equiv_{FC}$ asociado al factor de certeza viene dado por:

$$\equiv_{FC} (a, b, c, d) = \begin{cases} \dfrac{ad - bc}{(a+b)(b+d)} & \text{si ad > bc} \\ 0 & \text{si ad = bc} \\ \dfrac{ad - bc}{(a+b)(a+c)} & \text{si ad < bc} \end{cases} \tag{13}$$

Ahora en función de la definición anterior del cuantificador factor de certeza se procede a extenderlo al caso difuso. En este caso se considera a *A y B* como dos *itemsets* disjuntos difusos en $\tilde{D}$ y su tabla-4ft para cada nivel de restricción $\Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}$ que representa la unión de los *RL*-sets de $\widetilde{\Gamma}_A$ y $\widetilde{\Gamma}_B$. En función de lo anterior el cuantificador $\equiv FC$ extendido al caso difuso es:

$$\equiv_{FC} (A \rightarrow B) = \sum_{\alpha_i \in \Lambda_{\tilde{A}} \cup \Lambda_{\tilde{B}}} (\alpha_i - \alpha_{i+1})(\equiv_{FC} (a_i, b_i, c_i, d_i)) \tag{14}$$

D. Estrategia de Poda de Reglas

El uso de reglas de asociación difusa implica la posibilidad de obtener una base de reglas de gran tamaño lo cual tiende a aumentar con el número de transacciones.

Es necesario depurar el proceso de minado para acotar la cantidad de reglas a obtener lo cual puede disminuir también el tiempo de ejecución y el espacio para el almacenamiento de reglas.

Una de las formas de depuración es la poda de reglas, la cual consiste en una serie de criterios o técnicas que permiten filtrar una gran cantidad de reglas.

La poda se basa en ciertas propiedades o estructuras subyacentes de las reglas que permiten la aplicación de diversos métodos como los de agrupamiento de reglas, ontologías, hipergrafos, cobertura en función de la estructura de reglas, los de cobertura informativa, entre otros [25][26] [27].

Algunos mecanismos de poda, como es el caso de [27], destacan por su sencillez y facilidad de implementación. Este último se basa en la idea de calcular un conjunto de reglas más pequeño (conjunto de reglas informativo) pero que tiene la misma capacidad predictiva que la base de reglas obtenida por el minado convencional.

A continuación se resume la teoría relacionada con la obtención del conjunto de reglas informativo, mostrándose sólo aquellos que serán de utilidad para la poda de reglas del mecanismo propuesto [27]:

Definición 4. Dado un conjunto de reglas de asociación *R* y un itemset *P*, se dice que la predicción de *P* desde *R* es una secuencia de *ítems Q* la cual es generada al usar las reglas de R en orden descendente de confianza. Para cada regla *r* cuyo antecedente es un subconjunto de *P*, cada consecuente de *r* es agregado a *Q*. Después de agregar un consecuente a *Q*, todas las reglas con este consecuente se eliminan de *R*.

Para excluir aquellas reglas que nunca han sido usadas en la predicción se presenta la siguiente definición:

Definición 5. Sea R un conjunto de reglas de asociación y sea R^1 el conjunto de reglas de un solo consecuente en R. Un conjunto de reglas de asociación R_I es informativo sobre R si (1) $R_I \subset R^1$; (2) $\forall r \in R_I \nexists r' \in R_I$ tal que $r' \subset r$ y $conf(r') \geq conf(r)$; y (3) $\forall r'' \in R^1 - R_I, \exists r \in R_I$ tal que $r'' \supset r$ y $conf(r'') \leq conf(r)$.

Teorema 1. Sea R un conjunto de reglas de asociación. Luego el conjunto informativo de reglas R_I sobre R_A es el más pequeño subconjunto de R_A tal que, para cualquier *itemset* P, la secuencia de predicción de P desde R_I es igual a la secuencia de P desde R.

El teorema anterior expresa una propiedad muy importante del conjunto informativo de reglas, es decir, que su capacidad de predicción (de los consecuentes en la base de reglas) es igual al conjunto original de reglas pero es un subconjunto de este último y además es el más pequeño posible.

Lema 1. Si $sop(A) = sop(A \cup B)$, entonces para cualquier *Z*, la regla $XY \rightarrow Z$ y todas aquellas reglas más específicas no ocurren en el conjunto informativo de reglas.

El lema anterior permite descartar las reglas derivables las cuales no deben estar en el conjunto informativo de reglas. Se dice que una regla es derivable si su confianza y soporte pueden derivarse de otras reglas más generales. Específicamente, una regla $A \rightarrow B$ es derivable si hay un conjunto R de reglas, todas más generales que $A \rightarrow B$, tal que $A \rightarrow B$ y su soporte y confianza pueden ser obtenidos de R.

Lema 2. Si $sop(A \neg C) = sop(A \cup B \neg C)$, entonces la regla $A \cup B \rightarrow C$ y todas aquellas reglas más específicas no ocurren en el conjunto informativo de reglas. Donde $sop(A \neg C) = sop(A) - sop(A \cup C)$, es decir, el soporte de los *itemsets* que contienen a A pero no contienen a C.

E. Etiquetas Lingüísticas Basadas en el Contexto

En un trabajo anterior [28], se definió una extensión de consultas difusas sobre bases de datos relacionales en la que se interpretan términos vagos dependiendo del contexto. Aunque la definición de etiquetas difusas depende mucho de la preferencia del usuario, puede adecuar su interpretación a los datos reales, lo cual se conoce como el contexto de los datos.

Definición 6. [28] Un conjunto difuso es una generalización del concepto convencional de conjunto. Sea *U* un dominio específico, un conjunto difuso A es determinado por $\mu_A : X \rightarrow [0,1]$, llamada la función de membresía. Para cualquier $x \in U$, la medida $\mu_A(x)$ es conocida como el grado de membresía de x. Así, el conjunto difuso es definido como una colección de pares ordenados:

$$A = \{(x, \mu_A(x)) \mid x \in U\} \quad (15)$$

Definición 7. [28] Sea *U* un dominio numérico $x_1 \leq, x_2 \leq, x_3, \leq x_4 \in U$, se define la forma lineal de las funciones de membresía: $trapezoidal(x_1, x_2, x_3, x_4)$ como la función de membresía $\mu : X \rightarrow [0,1]$ dada en la Figura 1; se define también $left\ shoulder(x_2, x_3, x_4)$ como $trapezoidal(-\infty, x_2, x_3, x_4)$ y $right\ shoulder(x_1, x_2, x_3)$ como $trapezoidal(x_1, x_2, x_3, +\infty)$ (ver Figura 1).

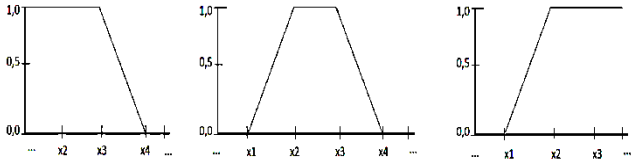

Figura 1: Forma Lineal de las Funciones de Membresía de Izquierda a Derecha: $left\ shoulder(x_2, x_3, x_4)$, $trapezoidal(x_1, x_2, x_3, x_4)$ y $right\ shoulder(x_1, x_2, x_3)$

Definición 8. [28] Un *Marco de Cognición* $\langle U, \mathrm{F}, \preceq \rangle$ es definido como una familia de conjuntos difusos F definido sobre el mismo universo del discurso *U*. El nivel de granularidad *K* de un *Marco de Cognición* es la cardinalidad de F. Resulta conveniente para cualquier aplicación establecer un orden total $\preceq$ sobre el marco de cognición. Se denota A_i al i-ésimo conjunto difuso en el *frame* F para $i \leq K$.

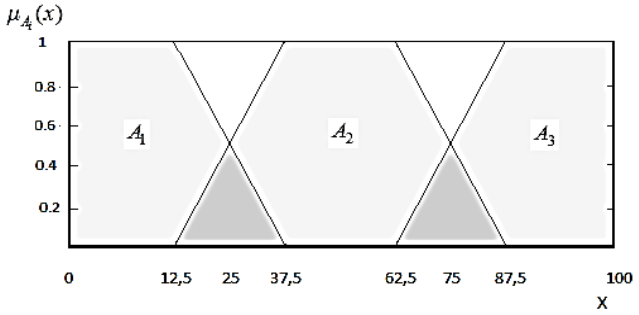

Figura 2: Ejemplo de Marco de Cognición con k=3

Definición 9. [28] Un Marco de Cognición es denominado una partición difusa si, y sólo si:

$$\forall A \in \mathrm{F}\ A \neq \varnothing \quad (16)$$

$$\bigcup_{A \in F} A = U \quad (17)$$

A continuación en la Tabla II se presentan los modelos teóricos para la interpretación de términos vagos basados en el contexto hasta k=3. Donde k es la granularidad de la categorización, *I* es el orden de la etiqueta y la tercera columna es la semántica de la i-ésima etiqueta. Además, cada P_q es el q-ésimo percentil para los valores del atributo.

Estos modelos representan las reglas semánticas para encontrar el significado de una etiqueta dependiente del contexto de unos datos específicos y usando los valores percentiles extraídos de dichos datos, según [28]. Estos percentiles permitirán inferir los términos vagos asociados a una etiqueta en términos de conjuntos difusos, los cuales serán expresados en función de las funciones de membresía difusa asociadas a los mismos.

Tabla II: Modelos Teóricos para la Interpretación de Términos Vagos Basados en el Contexto

K	I	Función de Membresía
2	1	$left\ shoulder(P_0, P_{37.5}, P_{62.5})$
	2	$right\ shoulder(P_{37.5}, P_{62.5}, P_{100})$
3	1	$left\ shoulder(P_0, P_{12.5}, P_{37.5})$
	2	$trapezoidal(P_{12.5}, P_{37.5}, P_{62.5}, P_{87.5})$
	3	$right\ shoulder(P_{62.5}, P_{87.5}, P_{100})$

III. MECANISMO PROPUESTO

En esta sección, se detalla la solución propuesta. En la sección A se describe el modelo planteado para tratar consultas en bases de datos incompletas. En la sección B, se describen las medidas de interés aplicadas por el componente de minería de reglas difusas perteneciente al mecanismo propuesto. En la sección C se describen los algoritmos principales que implementan el mecanismo propuesto.

A. Modelo de Tratamiento de Consultas

Aquí se plantea la aplicación de un método en bases de datos relacionales, donde no se realiza la imputación sino una estimación en forma dinámica en el momento de ejecutar una consulta que involucre un atributo que tenga valores nulos. De esta manera se evitan las desventajas antes descritas asociadas a la imputación de valores, antes descritas.

Por otro lado, se plantea el uso de reglas de asociación difusa, técnica ésta que permite obtener un conjunto de reglas tanto para atributos con valores continuos como con valores categóricos [29]. Este conjunto de reglas serán obtenidas sólo en función de los valores presentes en la base de datos y serán usadas para estimar los valores nulos que pudieran presentarse para los atributos que conforman la consulta. Estos valores estimados no reemplazarían los de la base de datos como en el caso de la imputación ni tampoco se hacen extensiones al modelo relacional como se realiza en los enfoques antes descritos, evitándose de esta forma las desventajas asociadas a la implementación de los mismos. También el usuario podrá conocer cuáles nulos pudiera arrojar su consulta y en qué casos se sustituyeron por estimaciones sólo a nivel de los resultados obtenidos de la consulta y en función de las preferencias expresadas por dicho usuario.

Con este planteamiento se pueden realizar consultas a la base de datos incompleta sin modificar su estado original y obteniendo un conjunto de respuestas más amplio gracias a la estimación de los valores nulos.

En particular el valor que se propone como estimado de un valor nulo sería una etiqueta difusa del atributo a estimar, que tiene de por sí una incertidumbre asociada a su representación y permite la adaptación a las preferencias particulares del usuario en cuanto a la precisión deseada para la estimación. Estas etiquetas son definidas combinado las preferencias del usuario y el contexto de los datos.

En cuanto al motor de inferencia utilizado para estimar los nulos, se cuenta con las reglas de asociación difusa que han probado ser de amplia aplicación y altamente configurables por parte del usuario. Cabe destacar que las reglas difusas deben ser generadas haciendo uso del motor de inferencia integrado. Una vez que la base de reglas es generada y almacenada en la base de datos por el motor de inferencia, es usada para realizar estimaciones en las consultas.

El modelo subyacente para minar las reglas de asociación difusa es un modelo maduro y extrapolable a muchos contextos y que permite una extensión natural de las medidas de interés de las reglas de asociación al caso difuso [23].

Otro aspecto importante en el algoritmo de minería de reglas de asociación difusa es el uso de técnicas de poda a fin de mantener acotada la cantidad de reglas y descubrir aquellas que son representativas de los patrones que se pueden derivar de la base de datos. Esto permite mejorar el uso del espacio de almacenamiento y reducir el tiempo de procesamiento con respecto a otros algoritmos de minería de reglas tradicionales.

Por otro lado, a diferencia de soluciones existentes anteriormente mencionadas, el mecanismo estimación de nulos aquí propuesto se integra completamente al sistema gestor de bases de datos, como se evidencia en la Figura 3.

Se propone como estimación del valor de un atributo faltante una etiqueta difusa. Para la estimación de dicha etiqueta, se consulta una base de reglas de asociación difusa. Esta base de reglas se obtiene al minar la base de datos del usuario y de acuerdo a las etiquetas difusas que este defina.

Como puede apreciarse en la Figura 3, tanto las etiquetas, la base de datos difusa y la base de reglas se guardan en tablas en forma permanente en la base de datos. De esta manera están a disposición del usuario y completamente integradas al sistema gestor de bases de datos.

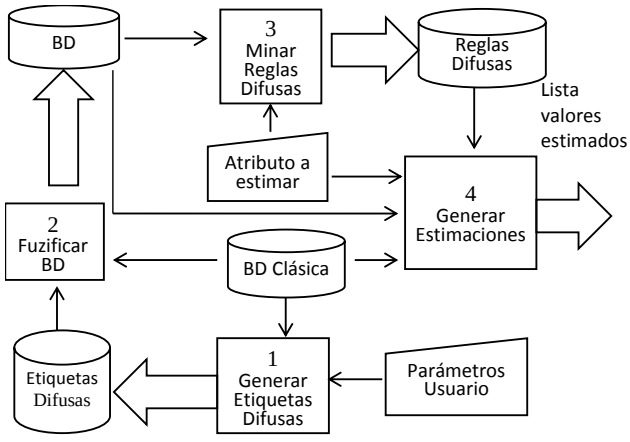

Figura 3: Arquitectura del Mecanismo de Tratamiento de Consultas en Bases de Datos Incompletas

B. Medidas de Interés del Modelo

Para el cálculo de la base de reglas difusas es necesario seleccionar un conjunto de medidas de interés adecuadas. Como se observó en la sección II, el modelo formal que se

propuso para la extracción de reglas difusas extiende de una manera natural las medidas de interés clásicas como lo son el soporte, la confianza y la certeza al caso difuso.

La ventaja de usar las medidas de interés extendidas al caso difuso reside en la capacidad expresiva que tiene la lógica difusa para representar el razonamiento humano. Por ejemplo, considere el valor *x* de un atributo *a* dentro de una transacción difusa. Suponga que dicho atributo *a* tiene un marco de cognición (partición difusa) asociado con dos etiquetas adyacentes o solapadas e_1 , e_2 (ver Figura 2). Además, suponga que el valor *x* del atributo se *a* ubica en el espacio donde se solapan e_1 y e_2. En este caso, el grado de pertenencia del valor *x* a la transacción se puede representar por dos g_1 y g_2, correspondientes a e_1 y e_2 respectivamente, estos grados son complementarios, es decir $g_1+g_2=1$. De este modo, cuando se calcule una medida de interés que involucre la etiqueta e_1 ésta se tomará en cuenta para el cálculo de la cardinalidad, si g_1 es mayor o igual al mínimo grado exigido por el usuario para la medida de interés (análogamente para e_2). Lo anterior tiene como ventajas una mayor flexibilidad para el cálculo de la cardinalidad, una mejor adaptación al razonamiento humano (el cual es gradual) y se minimiza la pérdida de información en el proceso de cálculo de dichas medidas.

C. Diseño de Algoritmos

En este apartado se describen los procesos y algoritmos principales que componen el mecanismo propuesto para el tratamiento de consultas en bases de datos incompletas.

El usuario define de acuerdo a sus preferencias, las etiquetas difusas que conforman el marco de cognición para los atributos que tendrán un tratamiento difuso. La interpretación de las etiquetas se hace basado en el contexto de la base de datos clásica *D* sobre la cual se quiere hacer la minería y las consultas. Siguiendo el modelo teórico [28] se generan los conjuntos difusos que definen cada etiqueta y se almacenan en una base de datos.

Con estas etiquetas lingüísticas se hace un proceso de fuzzificación. Esto es, se genera una base de datos difusa $\widetilde{D}$ asociada a *D*. $\widetilde{D}$ contiene los mismos registros que *D* pero representados en forma difusa. Para aquellos atributos que tiene un tratamiento difuso, en cada registro se sustituye el valor por la etiqueta correspondiente. En caso que el valor del atributo caiga en un intervalo de solapamiento, se tienen dos etiquetas e_1 y e_2 con sus respectivos grados g_1 y g_2.

Las etiquetas lingüísticas y la base de datos difusa se utilizan para minar las reglas de asociación difusa. En el caso que el usuario no introduzca etiquetas lingüísticas el sistema se comporta como un sistema preciso generando una base de reglas clásica.

Para cada atributo *a* del esquema de la base de datos *D* se genera un conjunto C_a de *itemsets* de cardinalidad 1. C_a contiene todos los *itemsets* de la forma $\{a=v\}$, siendo *v* un valor posible para *a* en la base de datos difusa $\widetilde{D}$.

Nótese que puede haber tanto valores precisos como difusos, dependiendo del tratamiento que el usuario quiera dar a cada atributo. En consecuencia, cuando se calcule el valor estimado de un atributo, si para este se definió un marco de cognición, el estimado será una etiqueta, de lo contrario, será un valor clásico.

A continuación se presentarán los algoritmos de minería de reglas de asociación difusa por niveles. Las reglas de nivel *k* son aquéllas que tienen *k* ítems en el antecedente. No todas las reglas son admisibles, sino aquéllas que cumplen con los niveles de calidad establecidos.

Algoritmo para minar de reglas difusas de nivel 1

Entradas:

$\widetilde{D}$ base de datos difusa asociada a *D*

b atributo en el esquema de *D* al cual estimar su valor

C_A conjunto de *itemsets* de cardinalidad 1 para atributos distintos de *b* , $C_A = \bigcup_{a \neq b} C_a$

C_b conjunto de *itemsets* de cardinalidad 1 para *b*

min_sop mínimo para la medida de soporte

min_conf mínimo para la medida de confianza

min_cert mínimo para la medida de confianza

Salida:

$\widetilde{R_1}$ conjunto de reglas de asociación difusa nivel 1 minadas a partir de $\widetilde{D}$

C'_A conjunto de *itemsets* frecuentes de cardinalidad 1 para atributos distintos de *b*

C'_b conjunto de *itemsets* frecuentes de cardinalidad 1 para el atributo *b*

Inicio

1) Inicializar C'_A en C_A y C'_b con C_b

   // todos los *itemsets* son potencialmente frecuentes

2) Para todo $i_A \in C'_A$ si $sop(i) < min\_sop$ entonces

   $C'_A := C'_A - \{i_A\}$ // el *itemset* no es frecuente)

3) Para todo $i_b \in C'_b$ si $sop(i) < min\_sop$ entonces

   $C'_b := C'_b - \{i_b\}$ // el *itemset* no es frecuente)

4) Inicializar $\widetilde{R_1}$ en $\varnothing$

5) Para todo $i_A \in C'_A$, $i_b \in C'_b$

   si (// la regla supera la calidad

   $sop(i_A \rightarrow i_b) \geq min\_sop \wedge$

   $conf(i_A \rightarrow i_b) \geq min\_conf \wedge$

   $cert(i_A \rightarrow i_b) \geq min\_cert$

) entonces $\widetilde{R_1} := \widetilde{R_1} \cup \{ i_A \rightarrow i_b \}$

Fin

Las reglas de nivel *k* se obtienen a partir de las de nivel *k*-1, es aquí donde se aplican los criterios de poda explicados en las bases teóricas de esta propuesta.

Algoritmo para minar de reglas difusas de nivel *k*

Entradas:

$\widetilde{D}$ base de datos difusa asociada a *D*

b atributo en el esquema de *D* al cual estimar su valor

C'_A conjunto de *itemsets* frecuentes de cardinalidad 1 para atributos distintos de *b*

C'_b conjunto de *itemsets* frecuentes de cardinalidad 1 para el atributo *b*

$\widetilde{R_{k-1}}$ conjunto de reglas de asociación difusa nivel *k*-1 minadas a partir de $\widetilde{D}$

min_sop mínimo para la medida de soporte

min_conf mínimo para la medida de confianza

min_cert mínimo para la medida de confianza

Salida:

$\widetilde{R_k}$ conjunto de reglas de asociación difusa nivel *k* minadas a partir de $\widetilde{D}$

Inicio

1) Inicializar $\widetilde{R_k}$ en $\varnothing$
2) Para todo $i \rightarrow i_b \in \widetilde{R_{k-1}}$, $i_A \in C'_A$

   si (// la regla supera la poda

   $sop(i) \neq sop(i \cup i_b) \wedge$

   $sop(i) - sop(i \cup i_b) \neq sop(i \cup i_A) - sop(i \cup i_A \cup i_b)$

)

   $\wedge$ (// la regla supera la calidad

   $sop(i \cup i_A \rightarrow i_b) \geq min\_sop \wedge$

   $conf(i \cup i_A \rightarrow i_b) \geq min\_conf \wedge$

   $cert(i \cup i_A \rightarrow i_b) \geq min\_cert$

) entonces $\widetilde{R_k} := \widetilde{R_k} \cup \{ i \cup i_A \rightarrow i_b \}$

Fin

El conjunto de reglas de asociación difusa minada a partir de la base de datos D se obtiene por la unión de las reglas de todos los niveles. El proceso de minería es iterativo y hasta que se llegue a un punto fijo, es decir, cuando se llegue a un nivel en que no se generen nuevas reglas. El conjunto de reglas obtenido se almacena en una base de datos a fin de poder ser usado al momento de una consulta.

A continuación se presenta el algoritmo para la estimación del valor de un atributo, En principio, este algoritmo se invocaría si el usuario desea obtener una respuesta estimada en lugar de un nulo.

Algoritmo para estimación de atributo

Entradas:

d registro en la instancia de una base de datos *D*

b atributo en el esquema de *D* al cual estimar su valor

$\widetilde{D}$ base de datos difusa asociada a *D*

$\tilde{R}$ base de reglas de asociación difusa minada de *D*

Salida:

v estimado para *b* en el registro *d*

Inicio

1) Recuperar la versión difusa $\tilde{d}$ de d en la base de datos $\widetilde{D}$.
2) Recuperar las reglas en $\tilde{R}$ con *b* en el consecuente, cuyo antecedente coincide con $\tilde{d}$
3) Ordenar las reglas recuperadas descendentemente por soporte, confianza y certeza
4) Elegir la primera en el orden, en caso de empate entre dos reglas se toma cualquiera de ellas
5) Seleccionar el consecuente de la regla elegida como *v*.

Fin

IV. Experimentos y Resultados

A. Diseño de Experimentos

Los datos elegidos para probar el sistema se encuentran explicados en [8], donde se estudia el uso de herramientas de aprendizaje automático para estimar dos parámetros de eficiencia energética de edificios residenciales. Todas las variables son numéricas dado que el mecanismo de estimación no es tan exigido en su rendimiento como cuando procesa atributos numéricos que posean etiquetas difusas. Esto se debe a la etapa previa de fuzzificar la base de datos, buscar las etiquetas difusas y evaluarlas durante una consulta.

El conjunto de datos se compone de los siguientes variables numéricas HL *(Heating Load*) y CL (*Cooling Load)*, en función de otras ocho variables de entrada, también numéricas: RC (*Relative Compactness*), SA *(Surface Area*), WA (*Wall Area)*, RA (*Roof Area*), OH (*Overall Height*), O (*Orientation*), GA (*Glazing Area)*, GAD *(Glazing Area Distribution*). En el presente análisis se usará la variable de salida HL y se le identificará como HL1, mientras que las variables de entrada se denominarán como RC1, SA2, WA3, RA4, OH5, O6, GA7 y GAD8.

En cuanto a la escogencia de los atributos que usarían etiquetas se realizaron una serie de pruebas combinando los atributos de a pares a fin de determinar el efecto de la interacción de los mismos en el rendimiento y precisión del algoritmo. De las pruebas realizadas se determinó preliminarmente que la combinación de los atributos RC1 y SA2 disminuía el rendimiento, debido quizás a que los atributos no son independientes [8], y debido a esto se eligieron las combinaciones RC1, WA3, RA4, GA7, HL1 y

SA2, WA3, RA4, GA7,HL1. Los atributos OH5, O6 y GAD8, pese a ser numéricos son considerados categóricos dada su baja granularidad y por lo cual no se definieron etiquetas para ellos.

Una vez elegidos los atributos el primer tipo de pruebas fue variando la cantidad de nulos y de etiquetas presentes en los atributos. En función de lo anterior se construyeron distintos escenarios para los cuales se obtuvieron distintas bases de reglas, las cuales a su vez se utilizaron para estimación. En la generación de las reglas se hicieron pruebas para medir la efectividad de los criterios de poda incorporados en el sistema.

El objetivo de estas pruebas es medir el impacto en el algoritmo en la medida que aumenta la cantidad de nulos o dicho de otra forma, en la medida que disminuye la información disponible. En cuanto al uso de etiquetas éstas permiten por un lado disminuir la granularidad de los atributos que las usan y por el otro al incrementarse su uso también aumenta la incertidumbre en los datos dada la naturaleza imprecisa de las mismas.

Adicionalmente a las pruebas preliminares para comprobar el funcionamiento del algoritmo, se realizaron unas pruebas de validación cruzada para comparar el mecanismo propuesto con otros dos [8]. Los dos mecanismos a comparar con el mecanismo de estimación son el método de los mínimos cuadrados iterativamente re ponderados o en inglés IRLS (*Iteratively Reweighted Least Squares*) y RF (*Random Forests*). La estructura de las pruebas es la descrita en [8] y consistió en repetir cien veces la ejecución del algoritmo. Para cada repetición fueron seleccionadas dos sub muestras, una del 10% de los datos totales como conjunto de prueba y la otra con el resto 90% como conjunto de datos de entrenamiento. Antes de escoger ambas sub muestras en cada repetición los datos totales son permutados. Finalmente en cada repetición se registra el MRE (*Mean Relative Error*), que luego se promedia para las cien repeticiones, y se especifica a continuación según [8]:

$$MRE = 100.\frac{1}{S}\sum_{i\in Q} | y_i - \hat{y}_i | / y_i \qquad (18)$$

Donde $\hat{y}_i$ es el valor estimado para el valor actual y_i de la variable HL1, perteneciente al valor i-ésimo del conjunto de entrenamiento o el de prueba. S representa el número de registros del conjunto de entrenamiento y Q son los índices de ese conjunto.

Cabe destacar que para probar el mecanismo propuesto se usa un *MRE* donde $|y_i - \tilde{y}_i| / y_i$, que representa el error relativo en cada estimación, es 0 si la etiqueta estimada corresponde a alguna etiqueta representada en la base de datos difusa y 1 en caso contrario. Por otro lado, dado que se tienen etiquetas lingüísticas definidas, el conjunto de registros con el valor de la etiqueta es un conjunto difuso. Debido a esto, S se calcula con la cardinalidad $\sum count$ de Zadeh [21], es decir se suman los grados de membresía de los elementos al conjunto. Así, si en una transacción hay una sola etiqueta esta suma 1, pero si hay dos etiquetas e_1 y e_2 con grados g_1 y g_2 respectivamente, la transacción suma g_1 para e_1 y g_2 para e_2. Por lo cual el valor de S es menor o igual al número de registros del conjunto de entrenamiento o el de prueba, dada la interpretación difusa que se le da a cada registro.

Por otra parte se realizaron 300 pruebas de validación cruzada: 100 para reglas con etiquetas sólo en el consecuente y 100 para cada una de las dos combinaciones de atributos con etiquetas en el antecedente y consecuente.

B. Resultados Obtenidos

Como puede observarse en la Tabla II y en la Figura 4 el porcentaje de estimación tiene una tendencia parecida independientemente de la presencia o no de valores nulos.

Esta tendencia se puede explicar ya que en cada momento el algoritmo construye la base de reglas considerando sólo la información disponible. Esta base de reglas posee una buena capacidad de generalización, ya que su poder predictivo no se ve afectado significativamente por la presencia de valores nulos en la muestra.

Por otro lado se puede apreciar una ligera disminución en el porcentaje de acierto promedio según la Tabla III cuando se utilizan etiquetas sólo en el consecuente de 94% versus un promedio de 89% cuando se usan etiquetas en antecedente y consecuente (ver último renglón), lo cual se puede confirmar en la Figura 4 y en el caso de usar valores nulos en distintas proporciones (ver Figura 5 y Figura 6).

Esta disminución pudiera deberse a que los atributos del antecedente no poseen una granularidad muy fina (2-12 valores diferentes), si se comparan con el atributo del consecuente (586 valores diferentes) [8], y al usar etiquetas difusas para ellos su granularidad se hace menos fina aún. Esto último genera una menor variedad de antecedentes y en consecuencia una menor variedad de reglas, disponiendo el algoritmo de menos oportunidades de generar una base de reglas de mayor calidad en cuanto a su poder predictivo.

Tabla III: Porcentaje de Aciertos %R Promedios por cada Tipo de Prueba para el Método Propuesto

Etiquetas Difusas / Nulos		Solo Cons	Antecedente y Consecuente			
			RC1,WA3, RA4,GA7	*SA2,WA, RA4,GA7*	*%R PROM*	
sin nulos		95	92	87	90	91
Solo consecentes	15%	96	93	88	92	89
	30%	94	89	85	89	
	45%	90	88	84	87	
Antecedentes y Consecuentes	15%	93	91	87	90	91
	30%	97	92	89	93	
	45%	95	92	87	91	
%R Promedio			91	87	90	
		94	89			

Otra razón que pudiera justificar una disminución en la precisión, es que al usar etiquetas tanto en el antecedente como en el consecuente se produce un aumento de la incertidumbre en la información necesaria para minado de la base de reglas, haciendo que ésta sea de menor calidad.

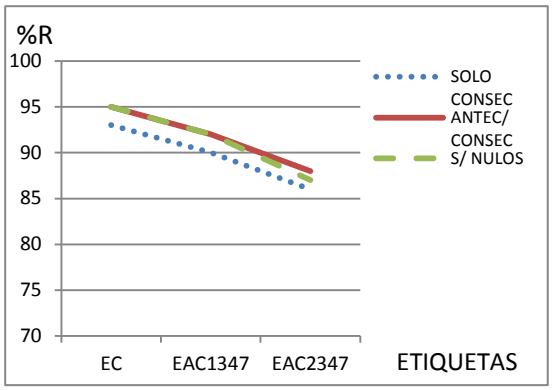

Figura 4: Porcentajes de Estimación para el Uso de Nulos para Etiquetas en el Consecuente y en el Antecedente para los Atributos 1347 y 2347

Por otro lado, el algoritmo podó en todos los casos por encima de un 50% de reglas del total de reglas calculadas (ver Figura 5 y Figura 6). Este porcentaje de poda resulta ser significativo, ya que se debe mencionar que al podar se previene de generar aquellas derivadas de las reglas podadas con lo cual la cantidad minada será lo más pequeña posible. De esta manera, el algoritmo propuesto reduce considerablemente tiempo de procesamiento y espacio de almacenamiento.

Adicionalmente, se observa una tendencia a disminuir el porcentaje de reglas podadas al aumentar el porcentaje de nulos (ver Figura 5 y Figura 6).

En primer lugar, en el caso de valores nulos sólo en el consecuente al aumentar los mismos esto trae como consecuencia una disminución de los valores que deben ser estimados (valores en el consecuente) y por ende la cantidad de reglas minadas tiende a disminuir lo que a su vez reduce el porcentaje de poda requerido como puede verse en la Figura 5. En esta misma figura se puede apreciar que el uso de etiquetas en el antecedente y en el consecuente redujo el porcentaje de poda requerido en contraste con sólo etiquetas en el consecuente, lo cual pareciera indicar un aumento en la calidad informativa de las reglas minadas por el uso de etiquetas.

En segundo lugar, en el caso de valores nulos en el antecedente y consecuente al aumentar los nulos también aumenta la incertidumbre, la cual en el caso de etiquetas en el antecedente y en el consecuente se incrementaría aún más. Pese a lo anterior se observa en la Figura 6 que los porcentajes de poda se mantuvieron bastante similares a los de la Figura 5. Esto pudiera indicar una gran estabilidad del componente de minado del sistema frente a niveles altos de incertidumbre en los datos.

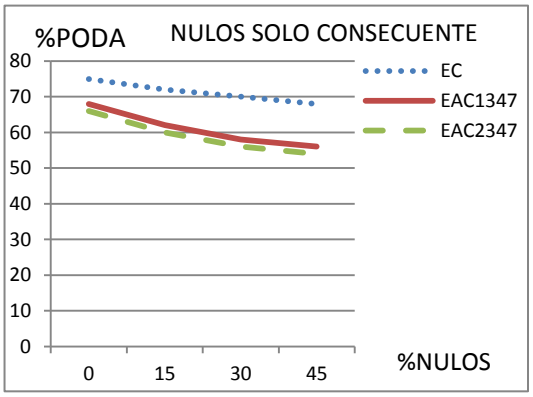

Figura 5: Porcentajes de Poda de Reglas para el Uso de Etiquetas en el Consecuente y en el Antecedente con Nulos sólo en el Consecuente

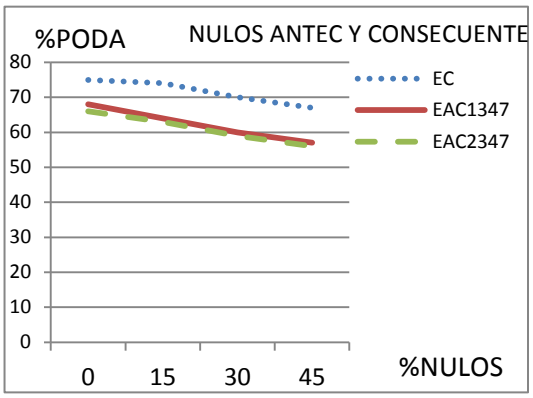

Figura 6: Porcentajes de Poda de Reglas para el Uso de Etiquetas en el Consecuente y en el Antecedente con Nulos en el Antecedente y Consecuente

Con respecto a las pruebas de validación cruzada (ver Tabla IV), el error promedio para los dos últimos casos que usan etiquetas en antecedente y consecuente es muy similar a diferencia de lo obtenido en las pruebas preliminares, lo cual puede indicar que no hay una diferencia apreciable entre usar el atributo 1 ó usar el 2 en cuanto a la precisión del algoritmo se refiere.

Por otro lado, se afianzó la tendencia obtenida en las pruebas preliminares en cuanto a que el error es menor usando sólo etiquetas en el consecuente a cuando se usan etiquetas en el antecedente y consecuente, obteniéndose en este último caso un error aproximadamente de 1 4/5.

Con respecto al método propuesto se puede observar de acuerdo a los resultados reflejados en la Tabla V que el error del mismo supera al del método RF pero es inferior a IRLS. Lo anterior se podría deber a que los datos usados para evaluar los métodos no son normales según [8], lo cual afecta a IRLS a diferencia del método propuesto que no requiere que los datos cumplan algún supuesto y por lo tanto la

precisión no se ve afectada. Además, pese a que la precisión es menor que RF, a diferencia de los otros dos métodos, tiene la flexibilidad de permitir modelar incertidumbre en todos los atributos usados como base para estimar y de acuerdo a las preferencias que el usuario de la base de datos ha preestablecido.

Tabla IV: Comparación de MRE entre las Pruebas Preliminares versus Validación Cruzada por cada Categoría de Prueba en cuanto al Uso de Etiquetas

	Uso de Etiquetas		
	Consecuente	***Anteced 1347-Consec***	***Anteced2347-Consec***
Prueba preliminar	5,45	7,79	12,6
Validación cruzada (error ± stdev)	7,39 ± 3,64	13,42 ± 6,59	13,14 ± 6,34

Por otro lado, la interpretabilidad de los estimados es más fácil en el método propuesto, debido a que se da como resultado una etiqueta lingüística la cual es más cercana al lenguaje natural. Esto permitiría, por ejemplo, que un experto en el contexto del problema pueda auditar con más facilidad la semántica de los patrones minados a fin de evaluar su fiabilidad.

El algoritmo, a diferencia de los otros dos métodos comparados, es bastante configurable en varios aspectos, como por ejemplo, en cuanto a los parámetros necesarios para su ejecución como lo son las medidas de interés. Éstas se les pueden asignar distintas combinaciones de valores a fin de experimentar distintos niveles de exigencia en cuanto a la precisión deseada. También se podría adaptar el algoritmo para no excederse de una cantidad máxima de reglas a ser minadas. Además, las etiquetas lingüísticas que usa el algoritmo se pueden definir siguiendo diversas estrategias de acuerdo a las preferencias del usuario y que en el caso particular de este trabajo se usó la definición de etiquetas basadas en el contexto [28].

El mecanismo propuesto tiene la versatilidad en cuanto a las características de los datos tratados, ya que trabaja con valores cuantitativos como categóricos en comparación con ILRS, además de soportar la presencia de nulos en los datos usados para realizar el proceso de minería en comparación con los otros dos métodos que no admiten nulos.

Otra ventaja resaltante con respecto a los otros dos métodos es que el método propuesto se encuentra encapsulado dentro del manejador de base de datos a través de una arquitectura de acoplamiento medio. Esta última característica le confiere una gran ventaja, con respecto a los otros métodos que se encuentran programados externamente a la base de datos, ya que se puede referenciar directamente desde la sintaxis SQL en las operaciones sobre los datos. Esto hace que la minería sea una extensión de las actividades habituales que un usuario especializado realiza dentro del manejador y aprovechándose además las ventajas de procesamiento, almacenamiento y rendimiento del manejador para la ejecución del algoritmo.

Por último, en cuanto al almacenamiento en la base de datos de las reglas minadas, esto es muy ventajoso debido a que no son requeridas nuevas ejecuciones del motor de inferencia. Lo anterior permite que el mecanismo propuesto pueda usarse para realizar tantas estimaciones como se requiera. En particular los otros dos métodos comparados, no disponen de la ventaja de almacenamiento permanente de su base de conocimientos en un manejador de base de datos.

Tabla V: Comparación de Método Propuesto con IRLS y RF

IRLS	**RF**	**Método Propuesto**
10,09 ± 1,01	2,18 ± 0,64	7,39± 3,64

V. Conclusiones y Trabajos Futuros

El objetivo fundamental del presente trabajo fue proveer un mecanismo automatizado que permitiera dar respuesta a consultas en base de datos relacionales en presencia de nulos, mediante la aplicación de reglas de asociación difusas que permitan estimar el valor de tales atributos. Este objetivo fue logrado a través de un modelo sencillo de tratamiento de consultas en bases de datos incompletas. En base a este modelo se construyó un mecanismo automatizado totalmente implementado en un manejador de base de datos a través del uso de funciones, procedimientos y estructuras de datos propias del manejador. Por otro lado, se definieron e implementaron como parte de este mecanismo las siguientes componentes:

Implementación de la estrategia automática de definición de etiquetas lingüísticas basadas en el contexto. Con esta estrategia el usuario de una manera sencilla puede dotar a su base de datos de una interpretación difusa más adaptada a la estructura de los datos sin necesidad de ser un experto en la semántica de los mismos. Implementación de un algoritmo de minería de reglas de asociación difusas basado en la extensión a un modelo formal para la representación y evaluación de reglas de asociación para la extracción de reglas difusas mediante niveles de restricción. Este modelo, además de ser sólido y probado a través de los años en diversos contextos, permitió darle de una manera bastante natural e intuitiva, una interpretación difusa a las medidas de interés clásicas usadas en la minería de reglas de asociación.

Un aspecto a resaltar del mecanismo es su aprovechamiento potencial a través de cualquier aplicación que haga consultas a bases de datos relacionales. Adicionalmente como valor agregado, se pueden explotar las bondades de escalabilidad, robustez y capacidades de administración propias de un SGBD y en éste está integrado el proceso de minería y no como un proceso externo, pudiéndose además guardar en la base de datos la base de reglas generadas para usos futuros. Además, una vez realizadas las diversas pruebas de rendimiento del algoritmo implementado se obtuvieron buenos resultados de estimación usando etiquetas difusas en diversos atributos, tanto en el antecedente como en el consecuente o atributo a estimar. Por otro lado, se comprobaron las bondades del algoritmo ante la presencia de nulos en la base de datos, los cuales no degradaron el

rendimiento en forma significativa del algoritmo. Adicionalmente, se comprobó experimentalmente la capacidad de éste para podar una cantidad significativa de reglas en base a los criterios de poda utilizados, los cuales reducen significativamente el tiempo de cómputo del algoritmo y el espacio de almacenamiento requerido. Finalmente se puede destacar que el mecanismo desarrollado no modifica la base de datos al realizar sus estimaciones evitándose así las desventajas asociadas a la imputación.

Entre los aspectos a desarrollar a futuro se pueden destacar:

Explorar la sensibilidad del algoritmo en cuanto al rendimiento se refiere y en cuanto a la calidad de la base de reglas generadas al realizar diferentes combinaciones para el soporte, confianza y certeza. Experimentar con otras medidas de interés investigadas en la literatura consultada a fin de que sean implementadas e integradas en el mecanismo desarrollado. De esta forma se podría probar el motor de inferencia implementado para la obtención de otro tipo de bases de reglas que sean útiles en otros contextos. Otro de los parámetros que se puede variar en cuanto al motor de minería, es explorar el uso de otras estrategias de definición de etiquetas distintas a las basadas en el contexto a fin de comparar la calidad de la base de reglas obtenida. También, se sugiere realizar diversas pruebas para estudiar el grado en el cual el motor de inferencia del mecanismo propuesto se ve afectado por la presencia de una menor o mayor cantidad de etiquetas difusas versus una menor o mayor proporción de nulos.

AGRADECIMIENTOS

Damos gracias a Aquél que aún en lo más difuso o incierto es capaz de proporcionarnos un orden: “pues Dios no es Dios de confusión, sino de paz.” (1 Corintios 14:33)

REFERENCIAS

[1] A. Motro, *Management of Uncertainty in Database Systems*, Book Modern database systems ACM Press/Addison-Wesley Publishing Co. New York, NY, USA, 1995.

[2] E. F. Codd, *Extending the Database Relational Model to Capture More Meaning*, ACM Transactions on Database Systems, vol. 4, no. 4, pp. 397-434, 1979.

[3] Z. Ma, W.J. Zhang, W.Y. Ma, *Extending the Relational Model to Deal with Probabilistic Data*, Journal of Computer Science and Technology, vol. 15, no. 3, pp. 230-240, May 2000.

[4] L.B. Othman, S.B. Yahia, *Yet Another Approach for Completing Missing Values*, CLA, vol. 4923 of Lecture Notes in Computer Science, pp. 155-169, Springer, 2006.

[5] L. Useche, D. Mesa, *Una Introducción a la Imputación de Valores Perdidos*, Terra Nueva Etapa, año/vol. XXII, no. 031, Universidad Central de Venezuela, Caracas, Venezuela, pp. 127-151, 2006.

[6] W.C. Hou, Z. Zhangl, N. Zhou, *Statistical Inference of Unknown Attribute Values in Databases*, in Proceedings of the second International Conference on Information and Knowledge Management, ACM, Washington, DC, USA, pp. 21-30, November 1993.

[7] J. Díaz, *Comparación entre Árboles de Regresión CART y Regresión Lineal*, Tesis de Maestría, Universidad Nacional de Colombia, Medellín, Colombia, 2012.

[8] A. Tsanas, A. Xifara, *Accurate Quantitative Estimation of Energy Performance of Residential Buildings Using Statistical Machine Learning Tools*, in Energy and Buildings, vol. 49, pp. 560-567, 2012.

[9] S.M. Chen, H.R. Hsiao, *A New Method to Estimate Null Values in Relational Database Systems Based on Automatic Clustering Techniques*, Information Sciences: an International Journal, vol. 169, no. 1, pp. 47–69, January 2005.

[10] J.W. Wang, C.H. Cheng, *An Efficient Method for Estimating Null Values in Relational Databases*, Knowledge and Information Systems: an International Journal, vol. 12, no. 3, pp. 379-394, Agosto 2007.

[11] K. Pandole, N. Bhargava, *Comparison and Evaluation for Grouping of Null Data in Database Based on K-Means and Genetic Algorithm*, International Journal of Computer Technology and Electronics Engineering (IJCTEE), vol. 2, no. 3, pp. 204-209, June 2012.

[12] S. Infante, J. Ortega, y F. Cedeño, *Estimación de Datos Faltantes en Estaciones Meteorológicas de Venezuela Vía un Modelo de Redes Neuronales*, Revista de Climatología 1578-8768, 2008.

[13] W. C. Beltran, H. Jaudoin, and O. Pivert, *Estimating Null Values in Relational Databases Using Analogical Proportions*, Information Processing and Management of Uncertainty in Knowledge-Based Systems, Springer International Publishing, Montpellier, France, pp.110-119, July 2014.

[14] K.N. ElSayed, *Estimating Null Values in Database Using CBR and Supervised Learning Classification*, International Journal of Advanced Computer Science and Applications (IJACSA), vol. 5, no. 6, June 2014.

[15] D. Otero García, *Imputación de Datos Faltantes en un Sistema de Información sobre Conductas de Riesgo*, Tesis en Master en Técnicas Estadísticas, Universidad de Santiago de Compostela, Galicia, España, 2011.

[16] A.P. Goicoechea, *Imputación Basada en Árboles de Clasificación*, Eustat Available in: http://www. eustat. es/documentos/datos/ct, vol. 4, 2002.

[17] F. Berzal, I. Blanco, D. Sánchez, and M. I. A. A. Vila, *Measuring the Accuracy and Interest of Association Rules*: *A New Framework*, Intelligent Data Analysis, vol. 6, no. 3, pp. 221-235, 2002.

[18] M. Delgado, N. Marín, D. Sánchez, and M.A. Vila, *Fuzzy Association Rules: General Model and Applications*, IEEE Transactions on fuzzy systems, vol. 11, no. 2, pp. 214-225, 2003.

[19] M. Kryszkiewicz, H. Rybinski, *Incomplete Database Issues for Representative Association Rules*, Lecture Notes in Computer Science, vol. 1609/1999, pp. 583-591, 1999.

[20] A.A. Chavan, V.K. Verma, *Treatment of Missing Values for Association Rules: A Recent Survey*, International Journal of Computer Applications, vol.70, no. 26, pp.1-4, May 2013.

[21] L. A. Zadeh, *Fuzzy Sets. Information and Control*, vol. 8, no. 3, pp. 338-353, 1965.

[22] C. Shyi-Ming, H. Chung-Ming, A New Approach to Generate Weighted Fuzzy Rules Using Genetic Algorithms for Estimating Null Values, Expert Systems with Applications, vol. 35, no. 3, pp. 905-917, 2008.

[23] M.D.R. Jiménez, *Modelado Formal para Representación y Evaluación de Reglas de Asociación*, Tesis Doctoral, Universidad de Granada, España, 2010.

[24] D. Sánchez, M. Delgado, and M. A. Vila, *A Restriction Level Approach to the Representation of Imprecise Properties*, in Proceeding of IPMU 2008, Málaga, Spain, pp. 153-159, June 2008.

[25] S. Kannan, R. Bhaskaran, *Association Rule Pruning Based on Interestingness Measures with Clustering*, International Journal of Computer Science Issues (IJCSI), vol. 6, no. 1, November 2009.

[26] S. Chawla, J. G. Davis, and G. Pandey, *On Local Pruning of Association Rules Using Directed Hypergraphs*, in Proceedings of the International Conference on Data Engineering (ICDE), Boston, MA, USA, vol. 4, pp. 832-841, Marzo-Abril 2004.

[27] J. Li, H. Shen, and R. Topor, *Mining Informative Rule Set for Prediction*, *Journal of Intelligent Information Systems*, vol. 22, no. 2, pp. 155-174, March 2004.

[28] C. Jiménez, H. Álvarez, and L. Tineo, *Context-Dependent Fuzzy Queries in SQLf*, On the Move to Meaningful Internet Systems: OTM 2012, Springer Berlin Heidelberg, pp. 763-779, 2012.

[29] A.Y. Rodríguez, J.F. González, J.F Martínez-Trinidad, J.A. Carrasco-Ochoa,; J. R. Shulcloper, *Minería de Reglas de Asociación sobre Datos Mezclados*, Reporte Técnico no. CCC-09-001, Coordinación de Ciencias Computacionales, INAOE, México, 2009.