Reglas de asociación en una Base de datos del área médica.
Association rules in a Database of medical area.

Ing. Agustín Sáenz López
Ingeniero Civil
Doctor en Ingeniería Civil, área: Sistemas de Planeación y Construcción.
Profesor-Investigador Facultad de Ingeniería, Ciencias y Arquitectura
Universidad Juárez del Estado de Durango. Gómez Palacio Durango, México.
Teléfono:. 871-7152017 E-mail: agusgpl@hotmail.com

Ing. Facundo Cortés Martínez
Ingeniero Civil Profesor Investigador
Doctor en Ingeniería con Especialidad en Sistemas de Planeación y Construcción
Facultad de Ingeniería, Ciciencias y Arquitectura de la Universidad Juárez del Estado de Durango. México.
Teléfono 871 7152017 E-mail: facundo_cm@yahoo.com.mx

Ing. Julio Roberto Betancourt Chávez
Ingeniero Civil
Doctor en Ingeniería Civil, área: Construcción Sustentable Profesor-Investigador Facultad de Ingeniería, Ciencias y Arquitectura Universidad Juárez del Estado de Durango. Gómez Palacio Durango, México.
Teléfono: 871-7152017 E-mail: jbetancourt@ujed.mx

Recibido: 15-03-17
Aceptado: 18-04-17

Resumen:

La minería de las reglas de asociación ha sido generalmente aplicada al área de las negocios en especial a las tiendas minoristas, en donde ha tenido un impacto muy importante en la extracción del conocimiento de las bases de datos que se generan en este tipo de negocios, en este artículo se pretende aplicar el algoritmo Apriori que es un algoritmo para la obtención de las reglas de asociación a una base de datos del área médica, en donde se muestra que este algoritmo también puede ser usado en áreas diferentes de los negocios y que las reglas de asociación obtenidas sirven para la toma de decisión.

Palabras Clave: Minería de datos, Reglas de asociación, Algoritmo Apriori, Confianza mínima,

Abstract:
The mining of association rules has generally been applied to the area of the business, in particular to the retail stores, where it has had a major impact on the removal of the knowledge of the databases that are generated in this type of business. The purpose of this article is to apply the Apriori algorithm that is an algorithm for the obtaining of the association rules to a database of the

medical area, where it is shown that this algorithm can also be used in different areas of business and that the association rules obtained are used for decision-making.

Keywords: Data Mining, Association Rules, Apriori Algorithm, Minimum Confidence,

Introducción:

La minería de las reglas de asociación es un área muy importante dentro de la minería de datos. Es un proceso no supervisado que tiene la finalidad de encontrar las reglas de asociación que se encuentran en las instancias de base de datos, su principal desarrollo ha sido en el área de los negocios, y en especial en los negocios minoristas, la base de datos que analiza la minería de las reglas de asociación son las transacciones que quedan registradas en una base de datos en las tiendas minoristas, en cada una de las transacciones el cliente lleva cierta cantidad de artículos de la canasta de artículos de venta que tiene ese negocio minorista, con esta información capturada, los algoritmos de minería de las reglas de asociación, lleva a cabo un análisis en donde determina cuales artículos se venden unos con otros, de esta asociación de artículos (ítem) se obtienen las reglas de asociación, debido a la gran cantidad de artículos que están tiendas tienen, existe una explosión de las reglas que se pueden generar, para obtener solamente las reglas más importantes se usan dos parámetros, el soporte mínimo y la confianza mínima que sirven para seleccionar cuales son las reglas de asociación más interesantes.

Una regla de asociación es una implicación de la forma $A \rightarrow B$, esta regla de asociación dice que cuando se compra el ítem A es probable que se compre el ítem B, tanto A como B pueden estar formados por uno o varios ítems. Otra interpretación que se le puede dar a la regla $A \rightarrow B$ es que cuando se cumple la condición A se lleva a cabo la acción B.

Trabajos anteriores:

En el trabajo de Mohammed 1997, se lleva un muestreo al azar de las transacciones de la base de datos para llevar a cabo la minería de las reglas de asociación. Con este muestreo se acelera el proceso de minería en más de un orden de magnitud y se reducen en forma dramática los costos del I/O, debido al muestreo existe una disminución del número de transacciones que serán consideradas para el análisis. Los patrones que se obtienen son representativos de la base de datos que se está analizando.

En el trabajo de Nayak 2001, se describe un algoritmo de reglas de asociación que busca reglas de asociación aproximadas. La aproximación ~AR permite que los datos que se ajustan a los patrones contribuyan al soporte de estos patrones. Esta aproximación es también útil en el procesamiento de los datos perdidos, ya que probabilísticamente contribuyen al soporte de los posibles patrones con los que tienen parecidos.

En el trabajo de Chen 2002, se forman conjuntos de reglas más pequeñas, es decir, un conjunto de reglas de asociación simples cada una teniendo en su consecuente un solo atributo. Este conjunto de reglas pueden ser usadas para obtener otras reglas de asociación, significando que el conjunto original de reglas basado en los algoritmos convencionales puede ser recuperado de las reglas simples sin perder información. El conjunto de reglas simples es mucho menor cuando se compara con el conjunto de todas las reglas. Además, en este trabajo se desarrollan los algoritmos que pueden manejar reglas del tipo $P \rightarrow ?$ or $? \rightarrow Q$.

En el trabajo de HAN 2004, se propone una estructura de árbol de patrones frecuentes, que es una extensión de la estructura de árbol fijo usada para el almacenamiento, la información importante de

los patrones frecuentes que son guardados en este árbol, de manera que la minería de los conjuntos de ítem de patrones frecuentes pueda ser minada de una manera más eficiente.

La eficiencia de la minería obtenida por este algoritmo es por medio de tres técnicas: (1) una gran base de datos es comprimida en una estructura de datos condensada y más pequeña en forma de árbol de manera que se evita continuos recorridos de la base de datos, (2) la minería basada en el arbol FP adopta un método de crecimiento de patrón fragmentado para evitar la generación de un numero grande de conjuntos de ítem candidatos (3) un método de divide y conquistaras es usado para descomponer las tareas de la minería conjuntos más pequeños de tareas, lo que reduce en forma dramática el espacio de búsqueda.

En el trabajo de Yen 2012, se propone el algoritmo SSR que tiene la finalidad de reducir el espacio de búsqueda e incrementar la velocidad del proceso de minería, el algoritmo está basado en la construcción de árboles de los patrones frecuentes. El algoritmo genera un sub-árbol para cada uno de los ítem frecuentes y después genera los candidatos en forma de sub-arboles. Para la generación de los conjuntos de ítem frecuentes candidatos el algoritmo solamente genera un pequeño conjunto de candidatos, y por lo tanto reduce significativamente el espacio de búsqueda. La estructura de almacenamiento del algoritmo SSR está basada en un árbol FP (FP-tree). El algoritmo SSR empieza recorriendo la base de datos de transacciones obtener los ítem frecuentes que cumplan con el condición del soporte mínimo. Después de generar todos los ítems frecuentes, el algoritmo SSR construye un árbol FP. Para construir el árbol FP (FP-tree) se ordenan las transacciones que solamente contienen los ítems frecuentes. Se construyen las ligas de los ítems para cada uno de los ítems frecuentes en la parte alta de la tabla y ayuda a buscar los nodos con el mismo item. Después de construir el árbol FP (FP-tree), se tienen dos pasos para cada uno de los ítems frecuentes.

El algoritmo Apriori:

El proceso del algoritmo Apriori empieza con la obtención de los llamados conjuntos de ítems frecuentes, los cuales son aquellos conjuntos formados por los ítems cuyo soporte obtenido de la base de datos es superior al soporte mínimo solicitado por el usuario. Debido al amplio uso del algoritmo Apriori, desde que se formalizó la inducción de reglas de asociación, la obtención de los conjuntos de ítems frecuentes es una tarea común en dichos algoritmos.

Agrawal 1994 en su algoritmo *Apriori* menciona que todo subconjunto de un conjunto de ítems frecuentes también será un conjunto de ítems frecuentes. Por lo tanto, el algoritmo *Apriori* obtiene en primer lugar los conjuntos de ítems frecuentes de tamaño 1 y, luego, los de tamaño 2 y así sucesivamente hasta que no se encuentren más conjuntos cuyos ítems no tengan el soporte mayor al soporte mínimo. Un ejemplo de cómo funciona el algoritmo *Apriori* es el siguiente, supongamos que tenemos una conjunto de transacciones en donde cada transacciones puede estar formada por uno o varios de los siguientes ítems; {a}, {b},{c},{d},{e}. Los conjuntos de ítems que el algoritmo buscara se muestran en la figura 1;

Los conjuntos de un solo ítem de la Figura son los obtenidos de una pasada en la base de datos y son los ítems cuyo soporte calculado en esa pasada es superior al soporte mínimo propuesto por el usuario.

Conjunto de ítems	Núm. De transacciones	
Conjunto de un solo ítem	$\{a\},\{b\},\{c\},\{d\},\{e\}$	
Conjuntos de dos ítems	$\{a,b\},\{a,c\},\{a,d\},\{a,e\}$ $\{b,c\},\{b,d\},\{b,e\}$ $\{c,d\},\{c,e\}$ $\{d,e\}$	

Conjuntos de tres ítems	$\{a,b,c\},\{a,b,d\},\{a,b,e\}$ $\{a,c,d\},\{a,c,e\}$ $\{a,d,e\}$ $\{b,c,d\},\{b,c,e\}$ $\{b,d,e\}$ $\{c,d,e\}$	
Conjuntos de cuatro ítems	$\{a,b,c,d\},\{a,b,c,e\}$ $\{a,b,d,e\}$ $\{a,c,d,e\}$ $\{b,c,d,e\}$	
Conjunto con 5 items	$\{a,b,c,d,e\}$	

Figura 1: Espacio de búsqueda del algoritmo A priori

Con esos conjuntos de un solo ítem se generan los conjuntos de dos ítems, para esto se combina el ítem $\{a\}$ con el ítem $\{b\}$ para formar el conjunto de dos ítems $\{a,b\}$, con este conjunto se calcula su soporte y si es mayor que el soporte mínimo antes definido, entonces este ítem de 2-item forma parte de los conjuntos de conjuntos de 2-item, ese proceso se realiza también para los ítem $\{a\}$ y $\{b\}$, y así sucesivamente.

Este proceso continua ahora con los conjuntos de tres ítems que se obtienen combinando los conjuntos de 1 item con los conjuntos de 2 items, un ejemplo seria combina el ítem $\{a\}$ del conjunto de 1 item con un conjunto de 2 items, por ejemplo $\{b,c\}$ para formar el conjunto $\{a,b,c\}$, este conjunto se buscado en la base de datos de transacciones, y si su soporte es mayor al soporte mínimo entonces forma parte del conjunto de tres ítems.

Considerando solamente el conjunto de tres ítems, las reglas de que podrían generar serían las siguientes:

$$\{a,b\} \rightarrow \{c\}$$
$$\{a,c\} \rightarrow \{b\}$$
$$\{b,c\} \rightarrow \{a\}$$

Por lo que este algoritmo generaría una gran cantidad de reglas de asociación.

De acuerdo con Neves 2008, se recomienda que, como parámetros de entrada del algoritmo, se defina un valor bajo para el soporte y un valor elevado para la confianza. De esta forma, en primer lugar se genera una gran cantidad de reglas y, posteriormente, se verifica la cohesión de las mismas a través de la medida de confianza. Una regla de asociación con un valor de confianza bajo no expresará un patrón de comportamiento en los datos y, por otra parte, un valor de soporte muy elevado probablemente llevaría a la perdida de patrones.

A pesar de ser muy utilizado actualmente, la ejecución del algoritmo *Apriori* es muy costosa, pues como se pudo observar anteriormente, el algoritmo genera muchas combinaciones de conjuntos de ítems y realiza posteriormente repetidas búsquedas por conjuntos de ítems frecuentes.

Parámetros de las reglas de asociación.

Los parámetros de las reglas de asociación sirven para medir que tan válidas y representativa son las reglas de asociación con respecto al conjunto de datos que se está analizando. El algoritmo *Apriori* genera una gran cantidad de reglas con muy poco ítems en la base de datos, de las cuales debemos de seleccionar cuales son las reglas más válidas y representativas a tomar en cuenta en base a los parámetros de las misma. Estos parámetros nos ayudaran a seleccionar de la gran

cantidad de reglas de asociación generadas por el algoritmo, dependiendo de la importancia de las mismas en base a los valores de estos parámetros.

Los principales parámetros de calidad de las reglas de asociación son el Soporte y la Confianza, la descripción de cada uno de ellos es la siguiente.

Soporte
El soporte de un item es la frecuencia con la cual este ítem se encuentra en las transacciones dividido entre el número de transacciones.

$$Soporte(A) = \frac{Numero\ de\ transacciones\ que\ contienen\ el\ item\ A}{Numero\ de\ transacciones\ de\ la\ base\ de\ datos}$$

Para obtener el soporte de una regla de decisión, por ejemplos $A \rightarrow B$, se obtiene con la siguiente ecuación

$$Soporte(A \rightarrow B) = \frac{Numero\ de\ transacciones\ que\ contienen\ los\ items\ A\ y\ B}{Numero\ de\ transacciones\ de\ la\ base\ de\ datos}$$

Confianza
La medida de confianza de una regla de decisión $(A \rightarrow B)$ es la división entre el soporte de la regla de decisión entre el soporte del antecedente de la regla de decisión, esto está representado por la siguiente ecuación:

$$conf(A \rightarrow B) = \frac{soporte(A, B)}{soporte(B)}$$

Metodología:

La base de datos que se usó para el análisis, es una conjunto de datos generada en las personas que fueron sometidas una operación quirúrgica y se encontraban en un lugar donde el medico tenía que decidir si los podía mandar a su casa, a un piso del hospital para recuperación o los mandaba a la sala de terapia intensiva, los médicos que decidían el destino de los post-operadores, lo hacían en base a 9 parámetros (atributos) que son los siguientes

1.- L-CORE (temperatura interna del paciente en grados C): high (> 37), mid (>= 36 and <= 37), low (< 36)

2.- L-SURF (temperatura superficial del paciente en grados C): high (> 36.5), mid (>= 36.5 and <= 35), low (< 35)

3.- L-O2 (saturación de oxígeno en %): excellent (>= 98), good (>= 90 and < 98), fair (>= 80 and < 90), poor (< 80)

4.- L-BP (última medición de la presión de la sanguínea): high (> 130/90), mid (<= 130/90 and >= 90/70), low (< 90/70)

5.- SURF-STBL (estabilidad de la temperatura superficial del paciente): stable, mod-stable, unstable

6.- CORE-STBL (estabilidad de la temperatura interna del paciente): stable, mod-stable, unstable

7.- BP-STBL (estabilidad de la presión sanguínea): stable, mod-stable, unstable

8.- CONFORT (confort percibido por el paciente, el valor de este atributo esta entre 0 y 20)

9.- DECISION (decisión a donde enviar al paciente):
I (el paciente es enviado a cuidados intensivos),
S (el paciente es enviado a casa),
A (el paciente es enviado a piso general del hospital para su recuperación)

En donde el último atributo es la clase y corresponde a la decisión que debe tomar el médico para mandar al paciente a cuidados intensivos, a su casa o a piso general en el hospital para recuperación.

La decisión es en base a la hipotermia es de gran interés después de la cirugía, los atributos corresponden rigurosamente a las mediciones de temperatura del cuerpo.

La clase puede tomar uno de los siguientes valores: I (el paciente es enviado a cuidados intensivos), S (el paciente es enviado a su casa), A (el paciente es enviado al piso del hospital general).

El total de instancias es de 87 para esta base de datos.

Experimentación y análisis de los resultados:

Para la experimentación se utilizó el Software WEKA (Waikato Environment for Knowledge Analysis) que es una colección de algoritmos de minería de datos y entre esos algoritmo se encuentra el algoritmo Apriori para la generación de las reglas de asociación. El software fue desarrollado por Universidad de Waikato, en Nueva Zelanda. Es un software de licencia libre.

Las características del algoritmo Apriori que se usó en WEKA son las siguientes; confianza mínima de 0.9, en base a esta condición encontró que el soporte mínimo que fue de 0.45 para las reglas de asociación generadas. El software encontró 11 conjuntos de ítems de un solo ítem, 12 conjuntos de ítems con 2 items y 3 conjuntos de ítems con 3 items, que cumplían con las condición de confianza mínima.

Las mejores reglas en base a la confianza se muestran en la tabla 2

Tabla 2: Las 10 reglas obtenida con el software WEKA

Regla de asociación	Parámetros de medición de la regla
1.- SURF-STBL=stable 44→CORE-STBL=stable 43	conf: (0.98) ; lift(1.05); lev(0.02) [2] ; conv (1.52)
2.- CONFORT=10 DECISION=A 48 → CORE-STBL=stable 46	conf: (0.96) ; lift(1.03); lev(0.02) [2] ; conv (1.1)
3.- L-CORE=mid COMFORT=10 44 → CORE-STBL=stable 42	conf: (0.95) ; lift(1.03); lev(0.01) [1] ; conv (1.01)
4.- DECISION=A 62 → CORE-STBL=stable 59	conf: (0.95) ; lift(1.02); lev(0.01) [1] ; conv (1.07)
5.- L-CORE=mid DECISION=A 41 → CORE-STBL=stable 39	conf: (0.95) ; lift(1.02); lev(0.01) [0] ; conv (0.94)
6.- CONFORT=10 65 → CORE-STBL=stable 61	conf: (0.94) ; lift(1.01); lev(0.01) [0] ; conv (0.9)
7.- L-SURF=mid 47 → CORE-STBL=stable 44	conf: (0.94) ; lift(1.01); lev(0) [0] ; conv (0.81)
8.- L-02=good 46 → CORE-STBL=stable 43	conf: (0.93) ; lift(1); lev(0.0) [0] ; conv (0.79)

9.- BP-STBL=stable 45 → L-CORE=stable 42	conf: (0.93) ; lift(1); lev(0.0) [0] ; conv (0.78)
10.- L-CORE=mid 57 → CORE-STABLE=stable 53	conf: (0.93) ; lift(1); lev(-0) [0] ; conv (0.79)

Tabla 3: Parámetros de calidad de las 10 reglas obtenidas

Regla	Soporte	Conf.	Lift	Leverage	Conviction
1	43	0.98	1.05	0.02 [2]	1.52
2	46	0.96	1.03	0.02 [2]	1.1
3	42	0.95	1.03	0.01 [1]	1.01
4	59	0.95	1.02	0.01 [1]	1.07
5	39	0.95	1.02	0.01 [0]	0.94
6	61	0.94	1.01	0.01 [0]	0.9
7	44	0.94	1.01	0 [0]	0.81
8	43	0.93	1	0 [0]	0.79
9	42	0.93	1	0 [0]	0.78
10	53	0.93	1	-0 [0]	0.79

De las 10 reglas que el software WEKA selecciono como mejores en base a la confianza, tenemos que todas ellas andan por debajo del 50% de soporte (support) mientras que la confianza está por arriba del 93% que es muy buen valor. Sin embargo debido a que el soporte está muy bajo estas reglas no se pueden considerar de calidad.

De las 10 reglas la mejor de ellas es la regla 6 que menciona

CONFORT=10 65 →CORE-STBL=stable 61

Esta regla menciona que cuando el paciente se encuentra en un confort de 10 su temperatura interna está estable, esta regla tiene una confianza del 0.94 y un soporte del 61%

En todas las 10 reglas se tiene una confianza del 0.93 al 0.98, que es un valor muy alto sin embargo el lift está apenas por arriba del 1, por lo que le quita calidad a estas reglas ya que existen más instancias en donde el antecedente y el consecuente aparecen separadas en diferentes instancias

La primera regla que es SURF-STBL=stable 44→CORE-STBL=stable 43
Presenta una confianza (0.98), un lift (1.05); un leverage (0.02) [2]; y un conviction (1.52)

Conclusiones:

En este trabajo se aplicó el algoritmo *Apriori* a una base de datos del área médica para obtener reglas de asociación, se usó el Software WEKA para el procesamiento de la información, teniendo como parámetro la confianza mínima que fue de 0.9, se obtuvieron las 10 primeras reglas que tenían la confianza mínima más alta y por arriba del 0.9, se obtuvieron reglas que servirían para la toma de decisiones en el ámbito de la salud, sin embargo las 10 reglas obtenidas adolecen todavía del problema que presentan estos algoritmos y es el de obtener reglas en algunas veces redundantes. Sin embargo se mostró que es posible obtener reglas de asociación en un campo diferentes que el de las tiendas minoristas.

Referencias:

AGRAWAL 1994; R. Agrawal and R. Srikant; "Fast algorithms for mining association rules," in *Proceedings of International Conference on Very Large Data Bases*, 1994, pp. 487-499.

CHEN 2002; Guoqing Chena,*, Qiang Weia, De Liub, Geert Wetsc; Simple association rules (SAR) and the SAR-based rule Discovery; Computers & Industrial Engineering 43 (2002) 721–733

HAN 2004; JIAWEI HAN,JIAN PEI,YIWEN YIN,RUNYING MAO;Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach; Data Mining and Knowledge Discovery, 8, 53–87, 2004

Mohammed 1997; M. J. Zaki, S. Parthasarathy, W. Lin and M. Ogihara. "Evaluation of sampling for data mining of association rules." Technical Report 617, University of Rochester, Rochester, NY, 1996.

Nayak 2001; Jyothsna R. Nayak and Diane J. Cook;Approximate Association Rule Mining; FLAIRS-01 Proceedings. 2001, AAAI (www.aaai.org).

Neves 2008; Inhauma Neves Ferraz, Ana Cristina Bicharra Garcia; Ontology In Association Rules Pre-Processing And Post-Processing; Pages-87-91, IADIS European Conference Data Mining, 2008

YEN 2012; SHOW-JANE YEN, CHIU-KUANG WANG1,LIANG-YUH OUYANG;A Search Space Reduced Algorithm for Mining Frequent Patterns; JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 28, 177-191 (2012)