Reglas de asociación para determinar factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas

Association rules for determining epidemiological risk factors of Chagas disease transmission

Marchán, Edgar[1,2] *; Salcedo, Juan[2]; Aza, Teresa[1]; Figuera, Lourdes[1]; Martínez de Pisón, Francisco[3] y Guillén, Pablo [2]

[1]Instituto de Investigaciones en Biomedicina y Ciencias Aplicadas (IIBCA), Universidad de Oriente, Cumaná, Venezuela
[2]Centro de Simulación y Modelos (CESIMO). Facultad de Ingeniería, Universidad de Los Andes. Mérida, Venezuela
[3]Grupo EDMANS, Universidad de La Rioja. Logroño La Rioja, España.
[*]emarchanmarcano@yahoo.es

Resumen

Este trabajo presenta una aplicación de la biblioteca ARules del paquete estadístico R, la cual permite encontrar asociaciones frecuentes entre las variables que constituyen una base de datos. El objetivo principal consiste en determinar los posibles factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas causada por Trypanosoma cruzi en poblaciones que reúnan las condiciones biogeográficas apropiadas. La base de datos contiene las características epidemiológicas más relevantes que están asociadas frecuentemente con la transmisión de la enfermedad: socioeconómicas destacando características de la vivienda, presencia de animales silvestres y domésticos que mantienen el ciclo de vida de T. cruzi, presencia de palmas, conocimiento de la enfermedad y diagnósticos serológicos. Mediante la aplicación de ARules se logra predecir y asociar en un 93% y 100% múltiples factores de riesgo para una serología positiva y negativa, respectivamente, mientras que por el método convencional de Chi-cuadrado se determinan sólo dos factores de riesgo asociados con la seropositividad. En vista de los resultados obtenidos, se concluye que la aplicación de reglas de asociación puede ser utilizada como una valiosa herramienta para establecer los factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas, constituyendo así la base para definir las políticas de salud por los organismos competentes orientadas a la prevención, control y vigilancia para erradicarla.

Palabras clave: Reglas de asociación, enfermedad de chagas, epidemiología, factores de riesgo, minería de datos.

Abstract

This work presents an application of the ARules library of R statistics package, which allow you to find frequent associations between the variables that constitute a database. The main objective is to determine the possible epidemiological risk factors of transmission of Chagas disease caused by Trypanosoma cruzi in population who possess the appropriate biogeographic conditions. The database contains the most relevant epidemiological characteristics that are often associated with the transmission of the disease: socioeconomic emphasizing characteristics of housing, presence of wild and domestic animals that maintain the life cycle of T. cruzi, the presence of palms, knowledge of disease and serologic diagnosis. By applying ARules is possible to predict and associate 93% and 100% of multiple risk factors for a positive serology and negative, respectively, whereas the conventional method of Chi-square determined only two risk factors associated with seropositivity. In view of the findings obtained we concluded that the application of association rules can be used as a valuable tool to establish epidemiological risk factors of Chagas disease transmission, thus forming the basis for defining health policies by agencies authorities aimed at the prevention, control and surveillance to eradicate it.

Key words: Association rules. chagas disease, epidemiology, risk factors, data mining.

1 Introducción

La enfermedad de Chagas o Tripanosomiasis americana es una afección causada por un protozoario flagelado, denominado Trypanosoma cruzi. El agente causal de la enfermedad fue descubierto por Carlos Chagas en 1907 en

Brasil, en muestras intestinales de insectos de la subfamilia Triatominae (chipos), que son los vectores de la enfermedad (Coura, 2007). La enfermedad de Chagas es un grave problema de salud pública en América Latina. Se estima que existen entre 15 a 16 millones de personas infectadas desde México hasta Argentina, 100 millones en riesgo de contraerla y 14000 muertes por año (DNDi 2010).

En Venezuela, la enfermedad de Chagas presenta un índice de prevalencia de 8.9% (Feliciangeli, 2009) y es endémica principalmente en zonas rurales de la mayor parte del territorio nacional (Añez y col, 1999), aunque en años recientes ocurrió un brote epidemiológico serio por transmisión oral en Escuela urbana de la capital del país (Garrido, 2007). La manifestación clínica predominante es la cardiopatía chagásica y fue responsable del 21% de las muertes asociadas a problemas cardíacos ocurridas en el país a mediados de esta década (Villalobos et al, 2004).

Las fuentes tradicionales de información sobre la epidemiología descriptiva de esta enfermedad, lesiones y factores de riesgo son generalmente incompletos, fragmentados y de incierta confiabilidad y comparabilidad. En este sentido, en el contexto de la minería de datos, las Reglas de Asociación son una herramienta valiosa y bien estudiada para el descubrimiento de relaciones de interés entre variables de una gran base de datos (Agrawal y col, 1993; Rajasethupathy y col, 2009).

Las reglas de asociación son una de las técnicas más conocidas dentro de la minería de datos. Este tipo de herramientas permiten encontrar relaciones frecuentes de aparición conjunta de objetos o ítems existentes en una base de datos. La gran ventaja de estas técnicas es que permite encontrar, dentro de un rango de medidas de significancia establecidas previamente por el analista, múltiples relaciones y mostrarlas en forma de reglas de conocimiento del tipo "SI … ENTONCES …". Este conocimiento puede ser analizado por los expertos del dominio con el objetivo de encontrar relaciones importantes, no triviales y desconocidas previamente.

Lógicamente, otra posibilidad, ante la gran cantidad de variables existentes, consiste en utilizar algoritmos de selección de variables que permitan la identificación de las más importantes y determinar la correlación de ellas con la variable de salida.

Generalmente, existen multitud de métodos de selección de variables que se dividen fundamentalmente en dos tipos (Liu and Yu, 2002; Peng et al., 2009; Blum and Langley, 1997): filter and wrapper. Los métodos filter operan independientemente, no utilizan ningún algoritmo de aprendizaje y realizan una evaluación según las características generales de los datos. Mientras que los wrapper requieren de algoritmos de aprendizaje predeterminados, lo que les da una mayor precisión aunque también conlleva un mayor coste computacional.

La ventaja de las reglas de asociación frente a todas estas técnicas es que no buscan correlaciones entre variables sino que buscan relaciones frecuentes entre ítems dentro de las variables de la base de datos. Esto significa que pueden existir ítems frecuentes entre variables que aparentemente no están correlacionadas, es decir, dentro de las variables algunos elementos pueden aparecer relacionados y con una cierta frecuencia, mientras que otros elementos no lo están. De este modo, es posible encontrar reglas que indiquen relaciones de dependencia entre ítems de varias variables relacionadas, en este caso, con la variable de salida.

Esta herramienta es muy fácil de utilizar pues solamente es necesario indicar los umbrales de significancia para obtener la lista de reglas que las cubran. La principal desventaja es que el número de reglas obtenidas puede ser muy elevado siendo, muchas de ellas, equivalentes, obvias o no útiles. Debido a esto, es necesario realizar un esfuerzo considerable para la interpretación, evaluación y selección de las reglas más interesantes. Además, solo funciona con variables categóricas por lo que es necesario discretizar aquellas variables que son numéricas. Por último, si el número de ítems diferentes es muy elevado y/o la base de datos es muy grande, el tiempo de cómputo puede ser considerable e incluso, algunas veces, inaceptable.Sobre la base de la compleja multifactorialidad que caracteriza a la enfermedad de Chagas, el presente trabajo tiene como objetivo determinar mediante reglas de asociación los factores de riesgo epidemiológico asociados con la transmisión de dicha enfermedad en una población rural del estado Sucre, Venezuela.

2 Materiales y métodos

2.1. Minería de datos

Se utilizó el proceso estándar de industria cruzada para minería de datos CRISP-DM, (Chapman y col, 2002)

2.2 Base de datos

Los datos que se utilizaron para establecer los factores de riesgo epidemiológico asociados con la enfermedad de Chagas, en las localidades rurales de la población de San Pedro, parroquia Santa Fe, municipio Sucre del estado Sucre, ubicada en la región Nororiental de Venezuela, fueron cedidos por el Laboratorio de Biología Molecular del Instituto de Investigaciones en Biomedicina y Ciencias Aplicadas de la Universidad de Oriente (IIBCA-UDO) y se recopilaron a través de una encuesta epidemiológica diseñada y validada por especialistas en sociología de la salud; aplicada a 293 familias para su caracterización epidemiológica. Adicionalmente, se contó con el diagnóstico serológico realizado aleatoriamente a 88 individuos, lo cual correspondió a una muestra estadísticamente representativa (30% del total de los individuos censados), en el cual se determinó la presencia (seropositividad) o ausencia (seronegatividad) de anticuerpos totales (IgM, IgA e IgG) anti-Trypanosoma cruzi, realizado en el Laboratorio de Fisiopatología del Instituto de Biomedicina de la Universidad Central de Vene-

zuela.

2.3 Descripción de las variables

Antes de transformar el banco de datos originales a la matriz binaria que lee el paquete estadístico R, para aplicar Arules, se eliminaron de la base de datos las características o atributos que se consideraron con información incompleta o redundante, de tal manera que se seleccionaron las características epidemiológicas relevantes que podrían constituir posibles factores de riesgo de transmisión de la enfermedad de Chagas, las cuales se estructuraron en:

Socioeconómicas:

Región: se reagrupó la variable sector en dos regiones: oriental y occidental, para favorecer la búsqueda de asociaciones específicas a cada región, ver Fig. 1.

Edad: discretizada en dos valores (<20 años transmisión activa y >21 años transmisión pasiva)

Sexo: representa el género del individuo

Ocupación: determina frecuencia de exposición de los individuos a riesgo de transmisión de la enfermedad.

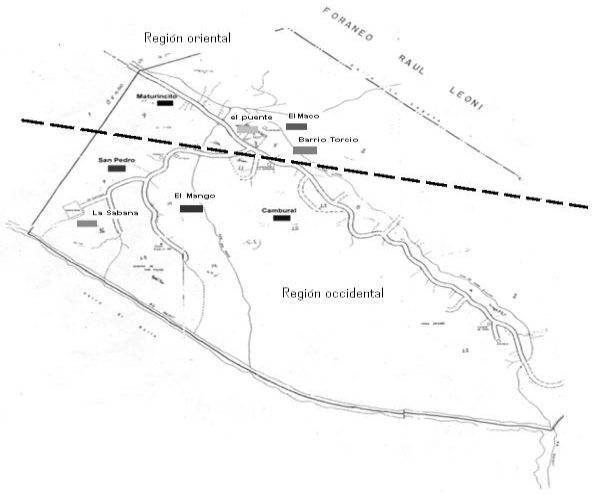

Fig. 1. Regiones estudiadas en la población de San Pedro

Características de la vivienda:

Tipo de vivienda: rancho, casa o quinta

Construcción de la vivienda: paredes, techo y piso utilizando materiales naturales propios de la zona que favorecen domiciliación del insecto vector o comerciales que la evitan.

Deposición de excretas: exposición del individuo a la picada del insecto vector, en baño con cloaca, con pozo séptico o al aire libre (mayor riesgo de contagio).

Tiempo en la zona: permite determinar si la enfermedad es endémica en la población en estudio, avalado por individuos seropositivos que nunca han visitado ni vivido en otras zonas con riesgo de transmisión.

Vivienda fumigada: frecuencia de rociamiento por organismos sanitarios competentes garantiza control de la población de insectos vectores en la zona.

Uso de insecticidas domésticos: periodicidad del uso de los mismos influye en una mejor protección ante el insecto vector.

Presencia de animales que mantienen el ciclo de vida de T. cruzi:

Silvestres o reservorios, los más frecuentes cachicamos y rabilpelados (zoonosis).

Domésticos relacionados con la transmisión, los más frecuentes: perros, gatos, cochinos, burros y aves de corral.

Presencia de palmas en el peridomicilio:

Hábitat por excelencia del insecto vector que favorece el establecimiento del ciclo de vida del parásito entre los animales domésticos y el hombre (antropozoonosis).

Conocimiento de la enfermedad:

Confirmación de existencia de la enfermedad de Chagas en la zona; es importante porque los prepara para conocer el modo de transmisión de la enfermedad y poder enfrentar así a sus enemigos naturales.

Conocimiento de insecto vector: determina el grado de prevención que puede implementar un individuo o la comunidad para evitar el riesgo de picadas por hematófagos infestados.

Contacto con el vector: establece presencia del insecto vector en la zona y fecha probable de contacto por picada con el individuo, permitiendo con ello posible entrada del parásito al torrente sanguíneo, convirtiéndolo en seropositivo.2.3 Reglas de asociación

Se utilizó la biblioteca ARules del paquete estadístico R para software libre, versión 0.6-5, (Hahsler y col, 2008).

Por definición:

Sea I = {i1, i2, . . . , in} un conjunto de n atributos binarios llamados ítems.

Sea D = {t1, t2, . . . , tm} un conjunto de transacciones llamados base de datos.

Cada transacción en D tiene una única transacción ID y contiene un subconjunto de ítems en I.

Una regla es definida como una implicación de la forma

$$X \Rightarrow Y, \text{ donde } X,Y \subseteq I \text{ y } X \cap Y = \phi \quad (1)$$

Medidas de significancia para seleccionar reglas de interés

$$Supp\ (X \Rightarrow Y) = Supp\ (X \cup Y) \quad (2)$$

[soporte]

$$Conf\ (X \Rightarrow Y) = Supp\ (X \cup Y) / Supp\ (X) \quad (3)$$

[confianza]

Una solución práctica al problema de encontrar muchas reglas de asociación que satisfagan los umbrales de cobertura y confianza, es filtrando el resultado usando medidas de interés adicionales como es el Lift:

$$Lift\ (X \Rightarrow Y) = Supp\ (X \cup Y) / (Supp\ (X)\ Supp\ (Y)) \quad (4)$$

Valores de Lift mayores que 1 indican que el consecuente es más frecuente en transacciones que contienen también el antecedente, que en transacciones que no la contienen. Ejemplo:

Análisis de la cesta de compras de un supermercado.

I = {carne, carbón, chorizo, pan}

y una base de datos que contiene 5 transacciones con los ítems que se muestran en la Tabla 1.

Tabla 1. Base de datos de compras en un supermercado

Transacción ID	Ítems
1	Carbón, carne, chorizo
2	Carbón, carne
3	Pan, chorizo
4	Carne, carbón, chorizo
5	Chorizo, carbón

Para la regla:
SI (carbón) ENTONCES carne

Supp (carbón) = 4/5 = 0,80

Conf (Si carbón entonces carne) = 0,6/0.8 = 0,75

Supp (carne) = 3/5 = 0,60

Lift (Si carbón entonces carne) = 0,75/0,6 = 1,25

Como contraste, consideramos otra asociación con la misma confianza:
SI (carbón) ENTONCES chorizo

Supp (chorizo) = 0.80

Lift (Si carbón entonces chorizo) = 0,94

Estos valores relativos de Lift indican que el carbón tiene una mayor asociación en la frecuencia de compra de la carne que en la de chorizo.

2.4 Algoritmo A priori

Se aplicó para la selección de los conjuntos de ítems que cumplen con un umbral de soporte, paso previo para generar las reglas de asociación que tengan un nivel de confianza mínimo Fig. 2.

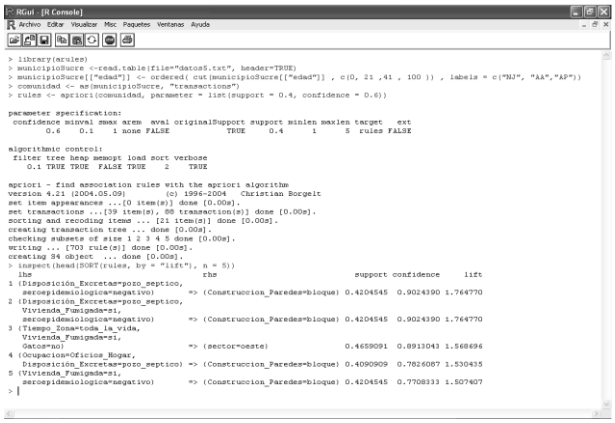

Fig. 2. Algoritmo Apriori.

3 Resultados

Previo a la aplicación del paquete ARules se realizó un análisis descriptivo para cada variable que conformó la base de datos bajo estudio. A modo de ilustración se presentan las más relevantes: la variable serología positiva de la enfermedad mostró que el 25% de la población ha presentado contacto con el parásito agente causal de la enfermedad.

Se observó también que el 50% de la población vive en casas de paredes de bahareque, las cuales favorecen la colonización del vector.

La variable presencia de los reservorios (rabipelados y cachicamos) indica que el 80% de los individuos viven en zonas donde habitan naturalmente los reservorios de T. cruzi.

Al aplicar el algoritmo Apriori estableciendo como umbrales de las medidas de significancia un soporte mínimo = 4% y confianza mínima = 60%, se encontraron un total de 225.428 reglas de asociación.

Debido a la gran cantidad de reglas obtenidas, se procedió a extraer los subconjuntos de reglas que contenían las variables epidemiológicas clásicas asociadas con los factores de riesgo epidemiológico de transmisión de la enfermedad de Chagas, encontrando:

Factores de riesgo epidemiológico asociados a la seropositividad ordenados en función del Lift

3526 reglas
Supp [4.55%, 29.54%]
Conf [60%, 93%]
Lift [1.51, 2.32]

SI (Ocupación = Oficios del Hogar, Perros = si, Rabipelados-Cahicamos = si y Palmas = si) ENTONCES serología = positiva. Supp = 13.64%, Conf = 92.31%, Lift = 2.320879

SI (Conoce la enfermedad = no, Vivienda fumigada = no, Perros = si y Rabipelados-Cahicamos = si) ENTON-

CES serología = positiva.

Supp = 13.63%, Conf = 85.71%, Lift = 2.155102

SI (Conoce la enfermedad = no, Vivienda fumigada = no, Aves de corral = si y Rabipelados-Cahicamos = si) ENTONCES serología=positiva.

Supp = 12.5%, Conf = 84.62%, Lift = 2.127473.

SI (edad = Joven, Ocupación = Estudiante, Construcción de paredes = bahareque y Tiempo en la zona = toda la vida) ENTONCES serología = positiva.

Supp = 11.36%, Conf = 83.33%, Lift = 2.095238.

SI (Construcción de paredes = bahareque, Tiempo en la zona = toda la vida, Vivienda fumigada = no, Perros = si) ENTONCES serología = positiva.

Supp = 10.23%, Conf = 81.82%, Lift = 2.057143.

Factores de riesgo epidemiológico asociados a la seronegatividad con máximo Lift

7914 reglas
Supp [4.55%, 60.23%]
Conf [60%, 100%]
Lift [0.99, 1.66]

SI (insecticidas = si, animales en la vivienda = si, Rabipelados-Cahicamos = no, y palmas = no) ENTONCES serología = negativa.

Supp = 11.36%, Conf = 100%, Lift = 1.660377.

SI (Construcción de paredes = bloque, palmas = no, Rabipelados-Cahicamos = no) ENTONCES serología = negativa, Supp = 11.36%, Conf = 100%, Lift = 1.660377.

SI (construcción de paredes = bloque, vivienda fumigada = si y Rabipelados-Cahicamos = si) ENTONCES serología = negativa.

Supp = 10.22%, Conf = 100%, Lift = 1.660377.

SI (Conoce la enfermedad = si, Vivienda fumigada = si, Perros = si y Construcción de paredes = bloque) ENTONCES serología = negativa.

Supp = 10.23%, Conf = 100%, Lift = 1.6603774.

Para efectos de comparación en la Tabla 2 se resumen las asociaciones de las variables epidemiológicas clásicas con la seropositividad encontradas aplicando el método convencional de Chi-cuadrado.

Tabla 2. Asociaciones de las variables epidemiológicas clásicas con la seropositividad aplicando el método convencional de Chi-cuadrado con un 95% de confianza, (Aza, 2003).

Variable epidemiológica	Significancia estadística
Animales domésticos relacionados con la transmisión de la enfermedad	***
Construcción de paredes con bahareque	***
Reservorios (rabipelados y cachicamos)	ns
Presencia de palmas	ns
Vivienda fumigada	ns
Ocupación	ns
Tiempo en la zona	ns
Conocimiento de la enfermedad	ns

*** altamente significativa; ns no significativa

4 Discusión

Considerando la seropositividad como parámetro determinante de la prevalencia de la enfermedad de Chagas, basados en la medida de significancia más potente LIFT, se pudo constatar que la población estudiada de San Pedro constituye un foco endémico de transmisión activa condicionada por la presencia de los siguientes factores de riesgo epidemiológico: reservorios (rabipelados y cachicamos), animales domésticos relacionados con la transmisión (perros y aves de corral), hábitat natural de los vectores (palmas), viviendas construidas con materiales naturales propios de la zona (paredes de bahareque) y desconocimiento de la enfermedad.

Asimismo, se pudo constatar que, frente al método de Chi-cuadrado utilizado convencionalmente en epidemiología, las reglas de asociación son capaces de evidenciar la presencia de relaciones ocultas entre algunos ítems de las variables que pudieran estar influenciando la transmisión de la enfermedad de Chagas en una población en estudio. De este modo, se puede afirmar que la utilización de reglas de asociación puede mejorar la obtención de conocimiento oculto frente al uso clásico de técnicas de selección de variables.

En contraste, se observó que los individuos que tienen viviendas construidas con paredes de bloques, usan insecticidas, tienen conocimiento de la enfermedad, ausencia de palmas peridomésticas aunque presentan reservorios y animales domésticos relacionados con la transmisión, se encontraron seronegativos, es decir, sin contacto hasta el momento del estudio, con el parásito Trypanosoma cruzi causante de la enfermedad de Chagas, a pesar de convivir en una comunidad con alto riesgo de contraerla.

5 Conclusiones

La aplicación de las Reglas de Asociación permite establecer de manera confiable, los factores de riesgo epide-

miológico presentes en una base de datos compleja que están asociadas con la transmisión de la enfermedad de Chagas en una población.

La comunidad de San Pedro es un foco endémico activo donde están presentes los factores de riesgo epidemiológico clásicos de transmisión de la enfermedad de Chagas.

El establecimiento de los factores de riesgo epidemiológico de la transmisión de la enfermedad de Chagas mediante Reglas de Asociación constituye una base sólida para la implementación de medidas de prevención, control y vigilancia epidemiológica por parte de los organismos sanitarios competentes.

6 Agradecimientos

A los Dres. Luis Briceño y Walter Mosca del Instituto de Biomedicina de la Universidad Central de Venezuela por la colaboración en el diagnóstico inmunológico. A la población de San Pedro por su valiosa participación para mejorar su calidad de vida.

Referencias

Agrawal R, Imielinski T y Swami A, 1993, Mining association rules between sets of items in large Databases. In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp. 207-216. ACM Press, URLhttp://doi.acm.org/10.1145/170035.170072. Fecha de consulta: el 17de noviembre de 2010.

Añez N, Carrasco H, Parada H, Crisante G, Rojas A, Fuenmayor C, González N, Percoco G, Borges R, Guevara P, y Ramírez J, 1999, Myocardial parasite persistence in chronic chagasic patients. Am. J. Trop. Med. Hug., 60:726-732.

Aza T, 2003, Evaluación seroepidemiológica del mal de Chagas en la población de San Pedro, parroquia de Santa Fe del municipio Sucre, estado Sucre. Trabajo Especial de Grado, Biblioteca Central UDO-Núcleo de Sucre. Pp. 50.

Blum, A.L. y Langley, P., 1997. Selection of relevant features and examples in machine learning. Artificial Intelligence 97(1-2), pp. 245-271.

Chapman P, Kerber R, Khabaza T, Reinartz T, Shearer C y Wirth R, 2002, CRISP-DM 1.0 Step by step data mining guide. http:/www.crisp.dm.org . Fecha de consulta: el 22 de noviembre de 2010.

Coura JR, 2007, Chagas disease: what is known and what is needed. A background article. Mem. Inst. Oswaldo Cruz. 102 (Supl1): pp. 113-122.

DNDi, 2010, Drugs for Neglected disease initiative, Chagas. http://www.dndi.org/diseases/chagas.html. Fecha de Consulta: el 17 de noviembre de 2010.

Feliciangeli DM, 2009, Control de la enfermedad de Chagas. Logros, Pasados y Retos Presentes. Interciencia, 34(6):393-399.

Garrido F, 2007. Vigilancia de la enfermedad de Chagas. En: Guía para el diagnóstico, manejo y tratamiento de la enfermedad de Chagas en fase aguda a nivel de los Establecimientos de Salud. M.P.P.S (eds). Venezuela. pp. 97.

Hahsler M, Buchta C, Gruen B y Hornik K, 2008, Mining association rules and Frequent Itemsets. The Arules package, version 0.6-5, 26/04/2008. http://R-Forge.R-project.org/projects/arules/ Fecha de consulta: el 17 de noviembre de 2010.

Liu H, y Yu L, 2002. Feature selection for data mining. Research Technical Report. Arizona State University.

Peng Y y Wu Z, Jiang J, 2009. A novel feature selection approach for biomedical data classification. Journal of Biomedical Informatics 43(1), 15-23.

Rajasethupathy K, Scime A, Rajasethupathy K S, Murray G, 2009, Finding ‘‘persistent rules”: Combining association and classification results. Expert Systems with Applications 36, pp. 6019–6024.

Villalobos L, De Sequeda, M y De Aponte, M. 1994. Enfermedad de Chagas: Transmisión vectorial y su control en Venezuela. Bol. Malariol. San. Amb., 34(1/4): 13 –21

Recibido: *27 de febrero de 2011*

Revisado: *10 de julio de 2011*