Extracción de reglas de asociación en una base de datos clínicos de pacientes con VIH/SIDA

P. Chausa Fernández[1], E.J. Gómez Aguilera[1], C. Cáceres Taladriz[1],

F García Alcaide[2], J.M. Gatell Artigas[2]

1 Grupo de Bioingeniería y Telemedicina, Universidad Politécnica de Madrid, Madrid, España, pchausa@gbt.tfo.upm.es

2 Servicio de Enfermedades Infecciosas, Hospital Clínic, Barcelona, España

Resumen

En la actualidad, las personas infectadas por el VIH con acceso a tratamiento retrasan indefinidamente su entrada en la fase SIDA de la enfermedad, convirtiéndose en pacientes crónicos. Un mayor conocimiento del comportamiento del virus y de cómo afecta a las personas infectadas podría conducirnos a optimizar el tratamiento y con ello mejorar la calidad de vida de los pacientes. En este contexto aparece la minería de datos, un conjunto de metodologías que, aplicadas a grandes bases de datos, nos permiten obtener información novedosa y potencialmente útil oculta en ellas. Este trabajo de investigación realiza una primera aproximación al problema mediante la búsqueda de asociaciones en una base de datos en la que se registran las historias clínicas electrónicas de personas infectadas que son tratadas en el Hospital Clínic de Barcelona.

1. Introducción

La aparición de los tratamientos antirretrovirales de alta eficacia marcó un punto de inflexión en la historia del SIDA disminuyendo drásticamente la mortalidad de las personas infectadas por el VIH. Así, en los países desarrollados, el SIDA es considerada una enfermedad crónica [1]. A lo largo de los últimos años la investigación médica ha conseguido tanto mejorar el tratamiento farmacológico como reducir los efectos secundarios y las resistencias que se producen en el organismo. La investigación principal se centra ahora en el desarrollo de vacunas terapéuticas y de prevención, vacunas con las que llegaríamos a un nuevo punto de inflexión en la evolución del SIDA. Según la opinión de la comunidad científica aún queda mucho por hacer antes de alcanzar dicho objetivo [2]. Mientras esperamos a que la investigación médica en vacunas dé sus frutos, debemos seguir trabajando en la optimización del cuidado de los pacientes mediante los recursos y la información de la que disponemos en estos momentos con el objetivo final de mejorar la calidad de vida de las personas que viven con el VIH/SIDA.

Un análisis profundo del proceso de la infección por VIH y la evolución de la enfermedad nos llevaría a la obtención de conocimiento oculto hasta ahora y potencialmente útil a la hora de tratar a los pacientes. Para realizar dicho análisis contamos con la existencia de historias clínicas electrónicas registradas en grandes bases de datos de los hospitales que tratan a las personas infectadas. Algunos hospitales tienen tal cantidad de datos registrados que el procesamiento de los mismos resulta extremadamente complejo dificultándose la extracción de la información que reside en ellos.

El análisis de una base de datos puede realizarse siguiendo varias estrategias. Un gran porcentaje de la base de datos contiene información extraíble mediante consultas SQL. Otra parte de menor tamaño requiere el uso de herramientas OLAP. Por último, hay una pequeña sección de cualquier base de datos que contiene información oculta sólo recuperable mediante minería de datos. La siguiente figura refleja esta situación [Figura 1].

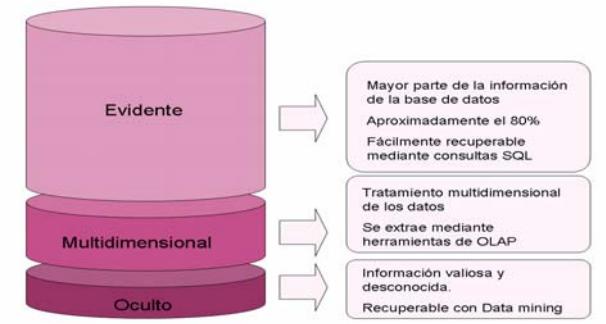

Figura 1. Obtención de la información en una base de datos

La minería de datos o data mining es un conjunto de metodologías y herramientas que permiten extraer el conocimiento útil oculto en un número elevado de datos. Es un campo multidisciplinar que se ha desarrollado en paralelo o como prolongación de otras áreas de conocimiento tales como la inteligencia artificial, la teoría de bases de datos o la estadística [3]. El "Knowledge Disovery in Databases" (KDD) es el proceso mediante el cual se aplica data mining a grandes bases de datos.

Existen diversas publicaciones que describen el uso de metodologías de data mining en bases de datos médicas para tareas de clasificación o búsqueda de asociaciones [4] [5]. En el campo del VIH/SIDA son masivamente utilizadas en el análisis de la generación de resistencias a los fármacos antirretrovirales [6].

2. Metodología: Reglas de asociación

Entre las metodologías que integra la minería de datos se encuentra la extracción de reglas de asociación, metodología que nos permite conocer la relación entre los

diferentes atributos de la base de datos que estemos analizando. Las reglas obtenidas mediante estas técnicas expresan patrones de comportamiento entre los datos en función de la aparición conjunta de valores de dos o más atributos.

2.1. Definiciones

Las reglas de asociación han sido el objetivo de muchos trabajos de investigación desde que Agrawal et al. propusieran el algoritmo de aprendizaje Apriori [7] y su utilización en grandes bases de datos [8]. Haciendo uso de su notación, podemos definir una regla de asociación como una implicación de la forma $X \Rightarrow Y$, donde X se denomina antecedente e Y consecuente. Tanto X como Y estarán formados por conjuntos de elementos pertenecientes a la tabla de transacciones T que estemos analizando. Una tabla de transacciones consta de un número indeterminado de registros que contienen diferentes secuencias de valores de los atributos que definen un registro. Los atributos que forman cada uno de los registros dependerán del campo de aplicación.

La búsqueda de asociaciones suele dar lugar a la obtención de un número muy elevado de reglas. Para seleccionar las más representativas el proceso debe ir seguido por una evaluación de las mismas. Las medidas más empleadas para estimar la validez de una regla son las que aparecen descritas a continuación:

- *Support* (cobertura): La cobertura expresa el tanto por ciento de registros de T que satisfacen la unión de los elementos del consecuente y del antecedente.
$$s(X \Rightarrow Y) = s(X \cup Y)$$
- *Confidence* (confianza): La confianza es una medida de la efectividad de una regla. Representa el porcentaje de casos en los que dado el antecedente se verifica la implicación.
$$c(X \Rightarrow Y) = s(X \Rightarrow Y)/s(X)$$
Puede utilizarse para estimar la probabilidad condicionada del consecuente dado el antecedente:
$$P(Y/X) = P(X \cup Y)/P(X) = c(X \Rightarrow Y)$$
- *Lift*: Cuantifica la relación existente entre X e Y:
  - *lift > 1*: X e Y positivamente correlados
  - *lift < 1*: X e Y negativamente correlados
  - *lift = 1*: X e Y independientes.
$$li(X \Rightarrow Y) = s(X \Rightarrow Y)/s(Y)$$

2.2. El algoritmo Apriori

El algoritmo Apriori tiene como objetivo la extracción de reglas de asociación de una base de datos de transacciones. Puede descomponerse en dos tareas:

1. Encontrar todos los conjuntos de elementos, que tienen una cobertura por encima de la mínima cobertura dada.
2. Utilizar los conjuntos de elementos con mayor cobertura de la fijada como umbral para generar reglas que superen un cierto nivel de confianza.

2.3. Proceso de extracción de reglas

La extracción de reglas de asociación es parte de un proceso más amplio que comienza con la preparación de los datos, que en la mayoría de los casos supone la mayor parte del esfuerzo, y termina con la representación del conocimiento adquirido. Este proceso iterativo aparece esquematizado en la siguiente figura [Figura 2].

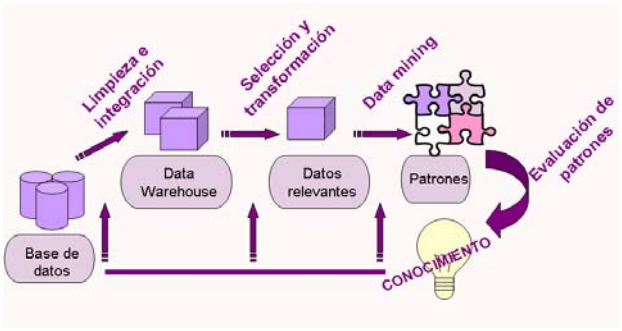

Figura 2. Proceso de extracción de conocimiento

Los pasos de los que se compone el proceso de extracción de conocimiento son los siguientes.

1. Limpieza de datos: se eliminan los datos irrelevantes y las posibles inconsistencias existentes.
2. Integración de datos: Se agrupan datos provenientes de fuentes diversas, en muchos casos heterogéneas.
3. Selección de datos: Se decide qué datos son relevantes para el análisis a realizar y se extraen de la base de datos.
4. Transformación de datos: Los datos se preparan en un formato adecuado para el proceso de data mining según la metodología elegida.
5. Data mining: Es el paso principal, el momento en que se aplican las técnicas que extraerán patrones potencialmente útiles. En nuestro caso dichos patrones serán reglas de asociación.
6. Evaluación de patrones: Se evalúan los resultados obtenidos utilizando medidas apropiadas en cada caso y se identifican los patrones de interés.
7. Representación de conocimiento: Se utilizan técnicas de visualización para que el usuario pueda comprender e interpretar los resultados.

3. Análisis de una base de datos clínicos

3.1. Descripción de la base de datos

Disponemos de una base de datos relacional que contiene información clínica y administrativa de pacientes infectados por el VIH y tratados en el Hospital Clínic de Barcelona desde la aparición de la enfermedad hasta Marzo de 2006. El primer paciente fue registrado en la base de datos en Julio de 1980.

La base de datos original está desarrollada en Microsoft Access 95 y para la realización de este trabajo de

investigación ha sido migrada a Mysql mediante Mysql Migration Toolkit, una herramienta de libre distribución [9]. Para la selección de los atributos con los que se va a trabajar y para una parte del preprocesado de la base de datos se ha utilizado Mysql Query Browser, también de libre distribución [9] La base de datos tiene un tamaño de 155MB y consta de 111 tablas que contienen información de 6277 pacientes.

3.2. Preprocesado de la base de datos: limpieza, integración, selección y transformación

El primer paso para poder trabajar con la base de datos es eliminar posibles inconsistencias, seleccionar los atributos con los que queremos trabajar, y preparar los datos de manera que la herramienta empleada para la extracción de reglas pueda trabajar con ellos.

La información se ha extraído de 6 tablas, vinculadas mediante el identificador de la historia clínica electrónica, valor unívoco para cada paciente. De cada una de estas tablas se han seleccionado aquellos atributos que a priori pueden ser significativos en la evolución de los pacientes. También se han creado atributos nuevos a partir de los originales, como, por ejemplo, el período en tratamiento de cada paciente.

La mayoría de los atributos seleccionados se han extraído de la tabla HISTORIA. Esta tabla registra información personal y determinados datos clínicos de los pacientes, como el sexo, el grupo de riesgo al que pertenecen o las fechas claves en la evolución de su enfermedad: fecha de serología (infección), de primer control, de SIDA o de exitus (muerte). De las otras 5 tablas se ha obtenido información sobre el número de visitas que ha realizado al especialista y al servicio de farmacia del hospital, la cantidad de infecciones oportunistas y neoplasias que ha padecido, la media del cumplimiento en la toma de la medicación y las veces que se le ha modificado el tratamiento. La tabla siguiente [Tabla 1] detalla los atributos seleccionados en cada una de las 6 tablas.

Tabla	**Atributos Seleccionados**
HISTORIA	SEXO, GRIESGO, EXITUS, TOXO, CMV, FNAC, FSERO, FPC, FSIDA, FEXITUS, TTO (Tiempo de tratamiento))
VISITAS	VISITAS (Número de visitas)
FARMACIA-VISITAS	VISITAS-FARMA, CUMP
IOS	IOS (Infecciones oportunistas))
NEOS	NEOS (Neoplasias)
TRATAMIENTOS	TTOS (Cambios de tratamiento)

Tabla 1. *Atributos de la tabla de transacciones*

Para disminuir la diversidad de valores de algunos atributos y poder extraer asociaciones más significativas, se han realizado agrupaciones dentro de los atributos seleccionados. En los de tipo nominal, la agrupación estaba definida de antemano. Es el caso del sexo (varón o mujer), grupo de riesgo (homosexual, bisexual, heterosexual, hemofílico, parenteral, consumidor de drogas por vía intravenosa o desconocido), exitus (si o no) o toxoplasmosis y citomegalovirus (positivo, negativo o desconocido). En otros casos ha sido necesario fijar umbrales para agrupar los posibles valores que pueden tomar los atributos. En el caso de las fechas se han agrupado por décadas. Para los atributos reales se han creado grupos basados en la incidencia de casos a lo largo del rango de variación. El resultado final es una tabla de transacciones con 6277 registros, uno por cada paciente, y 17 atributos para cada registro.

3.3. Búsqueda de reglas de asociación

Para proceder al análisis de los datos hemos utilizado la herramienta de software libre ARView [10], desarrollada específicamente para la extracción de reglas de asociación. Esta herramienta, programada en Java, facilita la visualización de las reglas obtenidas tras la aplicación del algoritmo Apriori y permite ordenarlas según diferentes parámetros: valores del consecuente, número de elementos, cobertura, confianza, valor de lift o métricas de validación adicionales. Dispone además de una interfaz para modificar los ficheros de entrada (datos a procesar) y salida (reglas generadas). Da la posibilidad de incorporar restricciones en la búsqueda, especificando si queremos que se tengan en cuenta o no ciertos valores de un atributo o en qué parte de la implicación nos gustaría que apareciesen. Estas restricciones se incorporan mediante un fichero de "aspecto" (*appearance file*). Podemos elegir la mínima confianza exigida y el rango de cobertura de las reglas generadas. También nos permite variar el número de elementos permitidos en la regla o dar indicaciones sobre el formato de entrada de los datos.

4. Resultados del análisis

Un primer análisis realizado con la configuración por defecto (confianza mínima del 80%, cobertura entre el 10% y el 100%, número de ítems entre 1 y 5) y sin imponer ninguna restricción, dio lugar a la generación de 14203 reglas de asociación. Para disminuir el número de reglas obtenidas y facilitar su análisis, variamos los parámetros e incorporamos restricciones que especifican en qué lado de la implicación deben situarse determinados atributos o que evitan la aparición en una misma regla de valores de atributos que aportan información similar. La siguiente tabla [Tabla 2] nos muestra la variación en el número de reglas obtenidas al modificar los valores por defecto.

Número de elementos	**Confianza**	**Cobertura**	**Nº Reglas**
1 min – 5 max	80%	10% - 100%	14203
2 min – 5 max	90%	10% - 100%	2148
2 min – 5 max	95%	10% - 100%	1504
2 min – 5 max	95%	20% - 100%	156
2 min – 7 max	95%	10% - 100%	2063

Tabla 2. *Número de reglas según el valor de los parámetros*

La tabla que aparece a continuación [Tabla 3] refleja algunas de las reglas obtenidas en las diferentes

realizaciones, junto con los valores de confianza, cobertura y lift asociados.

Regla	Conf.	Cob.	Lift
1. FSIDA-NULL, SEXO-V, TTOS-(0-10), NEOS-0 $\Longrightarrow$ EXITUS-NO	96,7%	27,8%	1,32
2. FSERO-(2000-2010), TTO-(1-10), IOS-0, NEOS-0 $\Longrightarrow$ EXITUS-NO	99,7%	13,8%	1,37
3. FEX-(1990-2000), IOS-(1-10), VISITAS-FARMA-0 $\Longrightarrow$ EXITUS-SI	100%	16,4%	4,01
4. GRIESGO-HMS, TOXO-POSITIVO $\Longrightarrow$ CMV-POSITIVO	85,8%	10,6%	1,41
5. CUMP-(60-80) $\Longrightarrow$ EXITUS-NO	98,9%	11,5%	1,35

Tabla 3. *Selección de reglas obtenidas y métricas asociadas*

5. Discusión

Este trabajo de investigación pretende aproximarse al problema de extracción de asociaciones y a la validación de una metodología concreta. Es en este contexto en el que deben analizarse los resultados obtenidos. Se han seleccionado las reglas que pueden ayudarnos a entender el significado de los valores de confianza, cobertura y lift. La primera nos muestra cómo los varones que no han entrado en la fase SIDA de la enfermedad, que no han sufrido muchos cambios de tratamientos y no han padecido neoplasias, no han fallecido. La número 3 tiene una confianza del 100% y un valor de lift muy elevado. Esto es obvio si nos fijamos en que aparecen valores de atributos directamente relacionados (fecha de exitus y exitus). También muestra la influencia de las infecciones oportunistas en la evolución de la enfermedad. Estas 2 reglas, como muchas de las obtenidas, reflejan información ya conocida, evidente en algunos casos.

La segunda puede hacernos pensar que las personas infectadas a partir de 1996, fecha en que aparecieron los tratamientos antirretrovirales, tendrán más probabilidades de evolucionar positivamente. La 4 indica una posible conexión entre ser homosexual, padecer toxoplasmosis y presentar citomegalovirus. La última regla demuestra la relación directa que existe entre la adherencia a los tratamientos y la correcta evolución de los pacientes.

6. Conclusiones y trabajos futuros

En este trabajo se describen los fundamentos teóricos y la aplicación de una metodología concreta para la búsqueda de asociaciones en una base de datos clínica de pacientes con VIH/SIDA. Comprobamos como la utilización de técnicas de minería de datos puede conducir a la extracción de patrones, confirmando en unos casos el conocimiento que se tiene acerca de la enfermedad y abriendo posibles vías de investigación médica para confirmar o descartar otras asociaciones obtenidas.

Todos los pasos del proceso son críticos para conseguir resultados significativos, destacando la selección y preparación de los datos a analizar. Es fundamental poseer un conocimiento profundo del campo de aplicación y contar con la colaboración de expertos para discriminar los datos que pueden ser relevantes y la forma en que se deben procesar para no perder información. Otro punto importante es el proceso posterior de validación. Este proceso requiere métricas y restricciones apropiadas que nos ayuden a descartar las reglas que no proporcionen información útil.

Siendo este un trabajo de aproximación al problema, el paso siguiente puede situarse en varios frentes. Cada etapa del proceso puede ser optimizada: la selección y preprocesado, la adaptación de los algoritmos al tipo de información empleada, la búsqueda de restricciones que reduzca el número de reglas y la selección de métricas más apropiadas. Sería interesante el análisis de diferentes herramientas software o el uso de otras metodologías para comparar y complementar los resultados. Todo ello con el objetivo último de obtener nuevo conocimiento médico que optimice el tratamiento y la calidad de vida de las personas que viven con VIH/SIDA.

Agradecimientos

Los autores expresan su agradecimiento al Servicio de Enfermedades Infecciosas del Hospital Clínic de Barcelona.

Este trabajo ha sido parcialmente financiado por el proyecto del MCYT TIC2002-02129, la Red de "Telemedicina" del MSC y por un convenio de colaboración con la empresa Glaxo Smith Kline.

Referencias

[1] Heighleyman L. Mortality trends: toward a new definition of AIDS?. *BETA*, vol 17, no 2, Winter 2005, pp 18-28 (ISSN: 1058-708X)

[2] International AIDS Vaccine Initiative. Developing and Delivering an AIDS Vaccine: Issues and Answers. Septiembre 2004. Actualizada en Junio de 2006.

[3] Hernández Orallo J, Ramírez Quintana MJ, Ferri Ramírez C. Introducción a la Minería de Datos. Pearson Educación SA, Madrid, 2004 (ISBN: 84-205-4091-9)

[4] Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller W, Muller R, Robson B, Apte C, Weiss S, Rigoutsos I, Platt D, Cohen S Knaus WA. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Computers in Biology and Medicine, Article in Press, December 2005 (ISSN: 0010-4825).

[5] Ordóñez C. Association rule discovery with the train and test approach for heart disease prediction. *IEEE Transactions on Information Technology in Biomedicine*, vol 10, no 2, Abril 2006, pp 334-343 (ISSN: 1089-7771).

[6] Draghici S, Potter B. Predicting HIV drug resistance with neural networks. *Boinformatics*, vol 19, no 1, 2003, pp 98-107 (ISSN: 1460-2059).

[7] Agrawal, R. and Srikant, R. Fast algorithms for mining association rules. Proc. 20th Int. Conf. Very Large Data Bases, (VLDB-94), 1994, pp 487-499.

[8] Agrawal, R., Imielinski, T., Swami, A. Mining association rules between sets of items in large databases. Proceedings of the International Conference on Management of Data (SIGMOD-1993), Washington, D.C., 1993, pp. 207-216 (ISBN:0-89791-592-5)

[9] Página web oficial de Mysql. http://www.mysql.com Consulta: Junio 2006.

[10] Página web de Christian Borgelt. http://fuzzy.cs.uni-magdeburg.de/~borgelt/index.html. Consulta: Junio 2006.