Extracción de reglas temporales complejas para la detección de fallos del tratamiento antirretroviral

P. Chausa Fernández[1,2], C. Cáceres Taladriz[1,2], F. García Alcaide[3], L. Sacchi[4], R. Bellazzi[4], E.J. Gómez Aguilera[2,1]

[1] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Madrid, España

[2] Grupo de Bioingeniería y Telemedicina, Universidad Politécnica de Madrid, Madrid, España, {pchausa, ccaceres, egomez}@gbt.tfo.upm.es

[3] Servicio de Enfermedades Infecciosas, Hospital Clínic, Barcelona, España

[4] Departamento de Informática y Sistemas, Universidad de Pavia, Pavia, Italia

Resumen

En la actualidad, las bases de datos clínicas contienen un gran volumen de información temporal que no está siendo suficientemente aprovechada y puede resultar fundamental para el óptimo cuidado de los pacientes. En este trabajo se describe un nuevo algoritmo que permite la asociación temporal del comportamiento de las variables que describen la evolución de los pacientes y la posterior obtención de reglas de interés clínico. Dicho interés es evaluado mediante el uso de diferentes métricas de demostrada utilidad en la extracción de conocimiento en bases de datos clínicas. Se presentan además los resultados obtenidos al aplicar este algoritmo a datos clínicos de pacientes con VIH/SIDA con objeto de detectar patrones de comportamiento de las variables que dan lugar a un fallo del tratamiento antirretroviral.

1. Introducción

Un nuevo informe presentado por ONUSIDA y la OMS expone que en 2007 33,2 millones de personas vivían con el VIH, otros 2,5 millones de personas se infectaron y 2,1 millones de personas fallecieron como consecuencia del SIDA [1]. Las previsiones para la situación de la epidemia en el futuro no son nada optimistas, presentando al VIH/SIDA en el 2030 como la tercera causa de muerte y la primera en DALYs (Disability Adjusted Life Years o años de salud perdidos) [2].

No cabe duda de que la aparición del tratamiento antirretroviral de alta eficacia en 1996 ha logrado frenar en muchos casos el avance de la infección y reducir con ello la mortalidad. Sin embargo, la CVRS (Calidad de Vida Relacionada con la Salud) en personas infectadas por VIH sigue siendo menor que la de la población general [3], incluso menor que la de personas con otras enfermedades como el cáncer o la depresión [4]. Esta calidad de vida se ve disminuida por la complejidad del tratamiento que en muchos casos debe ser interrumpido por su incapacidad de controlar el nivel de virus en sangre o por los efectos adversos asociados al mismo.

Un análisis profundo del proceso de la infección por VIH y de las causas por las que falla el tratamiento administrado, 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 los datos generados en las visitas que realizan las personas infectadas al hospital como parte del cercano seguimiento al que son sometidas. Algunos hospitales tienen tal cantidad de datos registrados que el procesamiento de los mismos resulta extremadamente complejo.

En este escenario aparece la extracción de conocimiento en bases de datos, metodología que ha demostrado su utilidad en el campo médico al utilizarse como método alternativo en la generación de hipótesis de investigación médica [5]. En el campo del VIH/SIDA son utilizadas de forma habitual en el análisis de la generación de resistencias a los fármacos antirretrovirales [6]. Podemos encontrar análisis de bases de datos de VIH mediante la adaptación de algoritmos estándar y su posterior aplicación a variables clínicas previamente seleccionadas y procesadas [7]. Se han desarrollado también metodologías que asocian temporalmente las mutaciones producidas en el gen del virus, los fármacos administrados a los pacientes y la evolución de los mismos [8]. Se demuestra así que la extracción de conocimiento en bases de datos puede utilizarse para descubrir patrones de comportamiento útiles en áreas específicas de investigación.

Este trabajo de investigación describe una metodología para la extracción de conocimiento en datos registrados en una base de datos clínica de VIH/SIDA. De forma más específica, el objetivo es detectar patrones, combinaciones de eventos que preceden a un fallo del tratamiento antirretroviral, diferenciando además cuando éste se produce por ineficacia, es decir, cuando la medicación no consigue controlar el nivel de carga viral en sangre o por toxicidad o efectos secundarios de los fármacos.

2. Materiales y métodos

El proceso de extracción de conocimiento utilizado consta de diferentes etapas. En una primera fase se realiza la selección, integración y preparación de las variables clínicas de estudio. Se obtienen después los patrones de interés y se establece la relación temporal entre los mismos, analizando qué combinaciones de patrones preceden a otro que se ha fijado como consecuente. En

este punto disponemos de una colección de reglas temporales complejas, entendiendo como complejidad el hecho de que el antecedente pueda estar formado por diferentes combinaciones de patrones. En la última etapa se extraen de esa colección las reglas que pueden ser consideradas de interés clínico haciendo uso de tres métricas de evaluación diferentes.

2.1. Preprocesado

El algoritmo diseñado se centra en los resultados de las analíticas de seguimiento que se realizan a las personas infectadas por VIH cada tres meses, aunque el procedimiento es generalizable a cualquier variable de características similares: aquellas que informan de una posible patología cuando su valor supera o es inferior a los límites de un intervalo de referencia concreto. Una vez extraídas las variables clínicas de interés de las diversas bases de datos e integradas en una nueva fuente de datos, se procede a codificar los valores en diferentes estados según sea la relación de éstos con su correspondiente intervalo de confianza. El límite inferior y superior de dichos intervalos así como otros posibles umbrales, son parámetros de diseño del algoritmo y serán establecidos por médicos especialistas.

2.2. Generación de reglas temporales

La obtención de los patrones considerados de interés por el usuario y el análisis de la relación temporal entre ellos utiliza el marco teórico conocido como "abstracciones temporales basadas en conocimiento (knowledge-based temporal abstractions)" [9].

De forma intuitiva, un patrón es un comportamiento o propiedad que puede distinguirse en una colección de datos. En datos temporales se suele asociar con el intervalo de tiempo en el que dicho comportamiento aparece. Cada uno de estos intervalos es lo que el marco teórico utilizado define como abstracción temporal básica. En nuestro caso nos centraremos en abstracciones temporales de tendencia, que son aquellas que representan patrones ascendentes, descendentes y estacionarios en series temporales de datos y todas las posibles combinaciones de los mismos. La primera tarea a realizar será agrupar la secuencia temporal de datos codificados resultantes de la fase de preprocesado en episodios o intervalos en los que se aprecie el comportamiento considerado de interés por el usuario. La figura siguiente (Figura 1) muestra un ejemplo de obtención de abstracciones temporales de tendencia a partir de una serie temporal de datos.

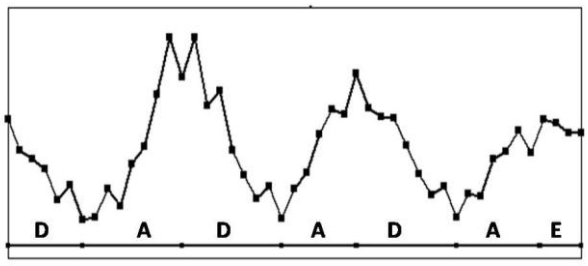

Figura 1. *Representación de abstracciones temporales básicas de tendencia*

En este trabajo se considera que se ha producido un evento ascendente cuando el valor de una variable ha cambiado de un estado a otro superior en un intervalo de tiempo inferior al máximo que se haya establecido como parámetro. Este valor máximo (INT_MAX) es otro de los parámetros de diseño del algoritmo y dependerá del ámbito de aplicación, es decir, de la variabilidad de los datos que se están manejando. Un valor elevado proporciona mayor número de episodios pero aumenta el riesgo de falsos positivos, es decir, de relacionar datos independientes entre sí. Un valor demasiado pequeño puede hacer que se pasen por alto algunas dependencias entre episodios. Como resultado de este paso obtendremos un conjunto de abstracciones temporales básicas en las que cada una de las variables incluidas en el estudio presenta él o los patrones de interés establecidos por el usuario.

El siguiente paso genera abstracciones temporales complejas haciendo uso del operador PRECEDES [10] que sintetiza 6 de los 13 operadores temporales definidos en el álgebra de Allen. Estos 6 operadores son: EQUALS BEFORE, FINISHES, OVERLAPS, MEETS y STARTS (Tabla 1).

EQUALS	BEFORE	FINISHES
|--A--|	|--A--|	|---A---|
|--C--|	|--C--|	|-C-|
OVERLAPS	**MEETS**	**STARTS**
|--A--|	|--A--|	|--A--|
|--C--|	|--C--|	|--C-----|

Tabla 1. *Operadores que sintetiza PRECEDES*

La distancia existente entre el antecedente y el consecuente se denomina GAP. Debemos establecer un valor máximo de GAP para poder considerar que dos episodios están relacionados. Este valor será un importante parámetro de diseño del algoritmo. PRECEDES analiza la relación temporal existente en las abstracciones temporales básicas y genera reglas de precedencia temporal en las que el antecedente puede estar formado por patrones tan complejos como el usuario defina.

2.3. Extracción de reglas: métricas de evaluación

El elevado número de reglas temporales que resultan de la etapa anterior hace necesaria la incorporación de métricas de evaluación con el objetivo de extraer las que sean más interesantes para los clínicos especialistas. Existe gran variedad de métricas que pueden utilizarse y la elección de una u otra dependerá del ámbito de estudio y del interés concreto del usuario final. Este trabajo utiliza tres de ellas: support, uncovered negative y accuracy. La definición original de estas métricas parte de la matriz de confusión, una herramienta de visualización de análisis predictivo formada por dos filas y dos columnas en la cual se representan el número de falsos positivos, falsos negativos, verdaderos positivos y verdaderos negativos.

Hechos		Predicción	
	Positivos	Negativos	Suma
Positivos	N_{TP}	N_{FN}	$N_{TP}+N_{FN}$
Negativos	N_{FP}	N_{TN}	$N_{FP}+N_{TN}$
Suma	$N_{TP}+N_{FP}$	$N_{FN}+N_{TN}$	N

Tabla 2. *Matriz de confusión que divide las muestras en verdaderos/falsos positivos y negativos*

Puesto que las métricas se van a emplear para medir el interés de reglas (A $\Rightarrow$ C), se hace necesario transformar la matriz de confusión anterior en otra que relacione las muestras con la aparición y no aparición de los episodios que forman el antecedente y el consecuente de cada regla.

Hechos		Predicción	
	A	~A	Suma
C	$N_{A\wedge C}$	$N_{\sim A\wedge C}$	N_C
~C	$N_{A\wedge\sim C}$	$N_{\sim A\wedge\sim C}$	$N_{\sim C}$
Suma	N_A	$N_{\sim A}$	N

Tabla 3. *Matriz de confusión que divide las muestras en aparición y no aparición de antecedente y consecuente*

A partir de estas dos tablas las métricas quedan definidas con las siguientes expresiones:

$$support = \frac{N_{TP}}{N_{TP}+N_{FP}+N_{TN}+N_{FN}} = \frac{N_{A\wedge C}}{N} = P(A\wedge C)$$

$$unc.neg = \frac{N_{TN}}{N_{TP}+N_{FP}+N_{TN}+N_{FN}} = \frac{N_{\sim A\wedge\sim C}}{N} = P(\sim A\wedge\sim C)$$

$$accuracy = support + uncovered\ negative$$

Tanto accuracy como uncovered negative han demostrado ser de gran utilidad a la hora de estimar el interés real de los especialistas clínicos [11]. El support es la métrica más común en procesos de extracción basados en el algoritmo original *Apriori* [12] y nos da una idea de la frecuencia con la que aparecen de forma conjunta el antecedente y el consecuente en el total de elementos a estudiar. En cambio, y tal y como se desprende de las ecuaciones, uncovered negative evalúa las muestras en las que no se da ni el antecedente ni el consecuente. El hecho de que el interés de los expertos se aproxime más a los resultados obtenidos mediante el uso de accuracy y uncovered negative en vez de support puede ser debido a que las reglas que éstas evalúan más favorablemente presentan menor riesgo.

3. Resultados

Los datos a analizar han sido proporcionados por el Servicio de Enfermedades Infecciosas del Hospital Clínic de Barcelona. Disponemos de la base de datos del propio servicio y de los resultados de las analíticas generales realizadas a los pacientes en los controles de seguimiento que tienen lugar cada 3 meses en el caso de los pacientes en situación estable. Los primeros datos fueron recogidos en septiembre de 1984 y se han incluido en el estudio todos aquellos anteriores a noviembre de 2007.

Se han seleccionado un total de 15 variables clínicas (Tabla 4). Tanto las variables incluidas en el estudio como los intervalos de referencia y demás parámetros de diseño han sido establecidos por los médicos especialistas del Hospital Clínic.

Glucosa	ASAT	Colesterol HDL
Urea - BUN	GGT	Trigliceridos
Acido Urico	Bilirrubina total	Amilasa
Creatinina	Colesterol total	Lipasa
ALAT	Colesterol LDL	Fosfata Alcalina

Tabla 4. *Variables clínicas incluidas en el estudio*

Este trabajo analiza ascensos en las variables que preceden a un fallo de tratamiento diferenciando cuando éste se produce por ineficacia o por toxicidad. Se han tenido en cuenta aquellos eventos producidos en los 6 meses anteriores al fallo. Como resultado se han obtenido dos conjuntos diferentes de reglas temporales a las que se ha evaluado utilizando las 3 métricas ya descritas. La siguiente tabla (Tabla 5) muestra algunos ejemplos de estas reglas.

A1	A2	A3	Accuracy	Un.Neg.	Support
Glucosa	Colesterol total	Triglic.	0,21393	0,21307	0,000863
Ácido Úrico	Bilirrubina	Lipasa	0,27601	0,276	1,37E-05
ALAT	ASAT		0,14059	0,1343	0,006283

Tabla 5. *Ejemplo de reglas temporales extraídas*

La Figura 2 nos permite evaluar la frecuencia de aparición de cada una de las 15 variables en las reglas que preceden un fallo de tratamiento. Si nos centramos en las 50 reglas con mayor valor de accuracy obtenemos los siguientes resultados.

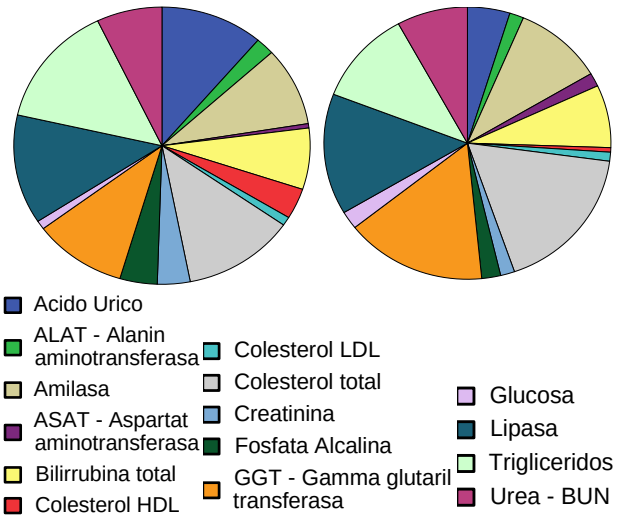

Figura 2.*Frecuencia de aparición de variables precediendo a un fallo de tratamiento por ineficacia y toxicidad*

Podemos hacer una comparación similar entre el primero de los gráficos y las 50 reglas con mayor support que preceden a un fallo por ineficacia (Figura 3).

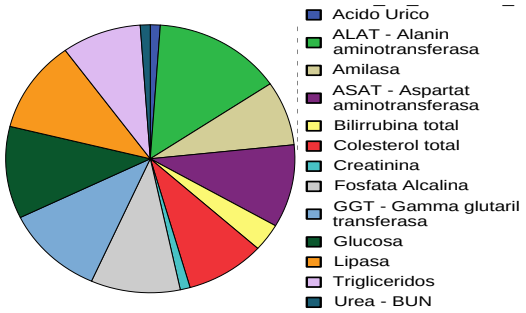

Figura 3.*Frecuencia de aparición de variables precediendo a un fallo de tratamiento por ineficacia*

4. Discusión

Los resultados obtenidos han sido analizados por los especialistas del Hospital Clínic. Algunas de las variables pueden asociarse, como la lipasa y la amilasa, marcadores del funcionamiento del páncreas, el GGT y la bilirrubina, del hígado o las grasas, representadas por el colesterol y los triglicéridos. Uniendo estas variables se obtienen tres grandes sectores relacionados directamente con el fallo de tratamiento tanto en el caso de fallo por ineficacia como por toxicidad. Esto se corresponde con la realidad, donde los marcadores del hígado y las grasas son los principales a la hora de realizar un cambio de tratamiento. La lipasa y amilasa, aunque también obligan a cambiar el tratamiento cuando presentan valores elevados, aparecen con menos frecuencia sobre todo en los últimos años de la epidemia. Podemos ver también una mayor influencia del ácido úrico, el colesterol y el GGT en el caso de fallos por toxicidad.

En el último gráfico se observan dos diferencias fundamentales. La primera es el aumento de la glucosa como precedente de un fallo de tratamiento y la segunda la aparición de dos marcadores íntimamente relacionados como son el ASAT y el ALAT, que nos indican el aumento de las transaminasas. Aunque hoy en día aparece con menos frecuencia, el aumento de glucosa solía ser un motivo frecuente de cambio de tratamiento en los primeros años de la epidemia. Tal y como dijimos anteriormente el ascenso del GGT y la bilirrubina obligan a realizar un cambio de tratamiento. La subida del ALAT y ASAT acompañando a éstos indican que el hígado ha sido seriamente afectado.

5. Conclusiones y trabajo futuro

La elección de un tratamiento antirretroviral óptimo es de vital importancia en la mejora de la calidad de vida de las personas que viven con VIH/SIDA. La detección de las causas por las cuales se producen fallos en el tratamiento pueden ayudar a los especialistas a personalizar y mejorar el cuidado de los personas infectadas. Este trabajo presenta el diseño de una metodología que permite el análisis temporal de los datos generados en el seguimiento de los pacientes. Los resultados obtenidos en este primer análisis han sido evaluados positivamente por especialistas clínicos, que también han mostrado su interés en la realización de estudios que permitan comparar patrones que preceden a fallos de tratamiento en los tres períodos principales de la epidemia: antes de la aparición de los tratamientos de alta eficacia (1980-1997), con tratamientos eficaces pero muy tóxicos (1998-2002) y con los tratamientos actuales (2003-2008). También han sugerido ampliar el tiempo de análisis al año anterior a la aparición del fallo e incorporar información sobre los fármacos administrados.

Referencias

[1] "Situación de la epidemia del SIDA: Diciembre 2007". ONUSIDA (Programa conjunto de las Naciones Unidas sobre VIH/SIDA). http://data.unaids.org/pub/EPISlides/2007/2007_epiupdate_es.pdf.

[2] Mathers, C.D.,Roncar, D. 2005. Updated projections of global mortality and burden of disease, 2002-2030: data sources, methods and results. Evidence and Information for Policy Working Paper. OMS.

[3] Miners, A.H., Sabin, C.A., Mocroft, A., Youle, M., Fisher, M., Johnson, M.. Health-related quality of life in individuals infected with HIV in the era of HAART. *HIV clinical trials*, vol 2, 2001, pp 484-492. (ISSN/ISBN: 1528-4336)

[4] Hays, R.D., Cunningham, W.E., Sherbourne, C.D., Wilson, I.B., Wu, A.W., Cleary, P.D., McCaffrey, D.F., Fleishman, J.A., Crystal, S., Collins, R., Eggan, F., Shapiro, M.F., Bozzette, S.A.. Health-related quality of life in patients with human immunodeficiency virus infection in the United States: results from the HIV Cost and Services Utilization Study. *The American Journal of Medicine*, vol 108, 2000, pp 714-722. (ISSN/ISBN: 0002-9343)

[5] Mullins, I.M., Siadaty, M.S., Lyman, J., Scully, K., Garrett, C.T., Miller, W.G., Muller, R., Robson, B., Apte, C., Weiss, S., Rigoutsos, I., Platt, D., Cohen, S., Knaus, W.A. 2006. Data mining and clinical data repositories: Insights from a 667,000 patient data set. Computers in biology and medicine 36, 1351-1377. ISSN/ISBN: 0010-4825.

[6] Draghici, S.,Potter, R.B. 2003. Predicting HIV drug resistance with neural networks. Bioinformatics 19, 98-107. ISSN/ISBN: 1367-4803

[7] Ramirez, J.C., Cook, D.J., Peterson, L.L., Peterson, D.M. 2000. Temporal pattern discovery in course-of-disease data. IEEE engineering in medicine and biology magazine : the quarterly magazine of the Engineering in Medicine & Biology Society 19, 63-71. ISSN/ISBN: 0739-5175

[8] Rashmi Raj, Martin J. O'Connor, Amar K. Das. An Ontology-Driven Method for Hierarchical Mining of Temporal Patterns: Application to HIV Drug Resistance Research. *AMIA 2007 Symposium Proceedings*. Chicago 2007, pp 614-619.

[9] L. Sacchi, C. Larizza, C. Combi, R. Bellazzi. Data Mining with Temporal Abstractions: Learning Rules from Time Series. *Data Mining and Knowledge Discovery*, vol 15, 2007, pp 217-247. (ISSN/ISBN: 1384-5810)

[10] Bellazzi R, Larizza C, Magni P, Bellazzi R (2005) Temporal data mining for the quality assessment of hemodialysis services. *Artificial Intelligence in Medicine*, vol 34, 2005, pp 25–39. (ISSN/ISBN: 0933-3657)

[11] Miho Ohsaki, Hidenao Abe, Shusaku Tsumoto, Hideto Yokoi, Takahira Yamaguchi. Evaluation of rule interestingness measures in medical knowledge discovery in databases. *Artificial Intelligence in Medicine*, vol 41, 2007, pp 177-196. (ISSN/ISBN: 0933-3657)

[12] Agrawal R, Srikant R. Mining sequential patterns. *Proceedings of the 11th international conference on data engineering*. Taipei, Taiwan, 1995, pp 3–14.