ACIMED
ISSN 1024-9435 *versión impresa*

ACIMED v.10 n.2 Ciudad de La Habana mar.-abr. 2002

Como citar este artículo

Aplicación de la minería de datos en la bioinformática

Juan Pedro Febles Rodríguez1 y Abel González Pérez2

Resumen

En los próximos años ocurrirá un avance espectacular de las ciencias biomédicas como resultado del proyecto Genoma Humano. Las nuevas tecnologías, basadas en la genética molecular y la informática, son claves para este desarrollo, pues ellas suministran potentes instrumentos para la obtención y el análisis de la información genética. La aparición de nuevas tecnologías ha posibilitado el desarrollo de la genómica, al facilitar el estudio de las interacciones de los genes y su influencia en el desarrollo de enfermedades, todo lo cual influye en el diagnóstico clínico, la investigación de nuevos fármacos, la epidemiología y la informática médica. En los últimos años, la minería de datos (data mining) ha experimentado un auge como soporte para las filosofías de la gestión de la información y el conocimiento, así como para el descubrimiento del significado que poseen los datos almacenados en grandes bancos. Esta permite explorar y analizar las bases de datos disponibles para ayudar a la toma de decisiones; además de facilitar la extracción de la información existente en los textos, así como crear sistemas inteligentes capaces de entenderlos, a esto se denomina comúnmente como minería de textos (text mining). Se describen sintéticamente los componentes básicos de la minería de datos y su aplicación en una emergente y trascendental actividad científica: la bioinformática.

DeCS: BIOLOGIA COMPUTACIONAL; INFORMATICA MEDICA; BASES DE DATOS; TOMA DE DECISIONES

El conocimiento es un recurso estratégico para el desarrollo económico y social contemporáneo. La información es el elemento básico principal en el proceso de adquisición, generación, gestión y trasmisión del conocimiento. Las tecnologías, métodos y herramientas asociadas con estos procesos se han desarrollado notablemente en los últimos años. La aparición de Internet ha facilitado compartir, en puntos distantes, los resultados científicos. Los análisis en línea (en inglés, *On-Line Analytical Processing, o OALP*), un enfoque novedoso, ha tomado gran fuerza en los últimos tiempos.

El aumento continuo de la disponibilidad de datos, en particular, a partir de las redes de comunicaciones y la aplicación de la computación de alto desempeño, con proezas como la descripción del genoma humano, convierten en imprescindible el empleo de técnicas y herramientas que le den sentido y utilidad a la información existente.

En los últimos años del presente siglo, ha alcanzado un auge la minería de datos, soporte de filosofías como la gestión de las relaciones de una organización con sus clientes. Su fin es explorar y analizar las bases de datos disponibles para ayudar a la toma de decisiones en las organizaciones, permite, a su vez, la extracción de la información existente en textos, así como crear sistemas inteligentes capaces de entenderlos, a lo que se le conoce, comúnmente, como minería de textos.

El surgimiento de técnicas como la minería de datos está asociado con la necesidad de procesar y analizar grandes volúmenes de datos, a fin de obtener información –mediante la consolidación de los datos- y

conocimientos, útiles a la toma de decisiones, y construir una experiencia, a partir de los millones de transacciones que registra una corporación en sus sistemas informáticos.1

El presente trabajo pretende sólo realizar una somera descripción de los componentes básicos de la minería de datos y su aplicación en una emergente y trascendental actividad científica: la bioinformática.

La minería de datos

La tecnología informática constituye la infraestructura fundamental de las grandes organizaciones y permite, hoy, registrar múltiples detalles de la vida de las empresas. Las bases de datos posibilitan almacenar cada transacción, así como otros muchos elementos que reflejan la interacción de la organización con otras organizaciones, clientes, o internamente, entre sus divisiones y empleados, etcétera.

Es imprescindible convertir los grandes volúmenes de datos existentes en experiencia, conocimiento y sabiduría, formas que atesora la humanidad para que sea útil a la toma de decisiones, especialmente en las grandes organizaciones y proyectos científicos. La búsqueda de información relevante siempre es útil a la administración empresarial: el control de la producción, el análisis de los mercados, el diseño en ingeniería y la exploración científica, porque pueden ofrecer las respuestas más apropiadas a las necesidades de información. Varias preguntas se relacionan frecuentemente con los datos, la información y el conocimiento. Su respuesta, demanda la participación de varios especialistas. ¿Cómo puede entenderse un fenómeno sobre la base de la interpretación de grandes volúmenes de datos? ¿De qué manera puede utilizarse la información para la toma de decisiones?, son algunos ejemplos de interrogantes comunes.

La respuesta a estas preguntas es el objetivo de la minería de datos, un conjunto de técnicas agrupadas con el fin de crear mecanismos adecuados de dirección, entre ellas puede citarse la estadística, el reconocimiento de patrones, la clasificación y la predicción.

Para descubrir patrones de relaciones útiles en un conjunto de datos se empezaron a utilizar métodos que fueron denominados de diferente forma. El término data mining, en inglés, no era, al principio, del agrado de muchos estadísticos, porque sus investigaciones estaban dirigidas a procesar y reprocesar suficientemente los datos, hasta que confirmasen o refutasen las hipótesis planteadas. Desde este ángulo, la minería de datos aplica una dinámica que se mueve en sentido contrario al método científico tradicional.

Con frecuencia, el investigador formula una hipótesis; luego, diseña un experimento para captar los datos necesarios y realizar los experimentos que confirmen o refuten la hipótesis planteada. Este es un proceso, que realizado de forma rigurosa, debe generar nuevos conocimientos.

En la minería de datos, por el contrario, se captan y procesan los datos con la esperanza de que de ellos surja una hipótesis apropiada. Se desea que los datos nos describan o indiquen el porqué presentan determinada configuración y comportamiento. Como afirma *Eduardo Morales*: “La más inocente mirada a los datos puede inspirar una hipótesis. Recuérdese que los humanos tienen un gran poder para generalizar e identificar patrones. Luego entonces, validar una hipótesis inspirada por los datos en los datos mismos, será numéricamente significativa, pero experimentalmente inválida.” 2

No es ocioso insistir, en que las técnicas de minería de datos no pueden utilizarse para confirmar o rechazar hipótesis, porque puede conducir a errores fatales. Su función es otra, como antes se expresó, se trata de explorar datos, darles sentido, convertir un volumen de datos, que poco o nada aportan a la descripción, en información para interpretar un fenómeno, para adoptar decisiones de acuerdo con las necesidades.

Componentes de la minería de datos

Las componentes básicas de los métodos de la minería de datos son:

1. *Lenguaje de representación del modelo:* comprende las suposiciones y restricciones utilizadas en la representación empleada.
2. *Evaluación del modelo:* incluye el uso de técnicas de validación cruzada para la predictividad y

aplicación de principios como el de máxima verosimilitud o el de descripción mínima para evaluar la calidad descriptiva del modelo.
3. *Método de búsqueda:* puede dividirse en búsqueda de parámetros y del modelo, determinan los criterios que se siguen para encontrar los modelos.

Algunas de las técnicas más comunes usadas en la minería de datos son:

- Árboles de decisión y reglas de clasificación.
- Métodos de clasificación y regresiones no-lineales.
- Métodos basados en ejemplos prototípicos.
- Modelos gráficos de dependencias probabilísticas.
- Modelos relacionales.

La minería de datos y el descubrimiento de conocimientos en bases de datos

Existe cierta tendencia a identificar como sinónimos a la minería de datos y el descubrimiento de conocimientos en bases de datos, que de forma abreviada se refiere con las siglas KDD (del inglés *Knowledge Discovery in Data Bases*), la convergencia del aprendizaje automático, la estadística, el reconocimiento de patrones, la inteligencia artificial, las bases de datos, la visualización de datos, los sistemas para el apoyo a la toma de decisiones, la recuperación de información y otros muchos campos.

El KDD es el proceso completo de extracción de conocimientos, no trivial, previamente desconocidos y potencialmente útil a partir de un conjunto de datos, mientras que «la minería de datos es una compilación de técnicas reunidas para crear mecanismos adecuados para la toma de decisiones. Entre estas técnicas se pueden citar la estadística, el reconocimiento de patrones, la clasificación y la predicción, la excavación de información relevante de la administración empresarial, el control de la producción, el análisis de los mercados, el diseño en ingeniería y la exploración científica.” 3 En otras palabras, el concepto minería de datos se asocia al proceso de construcción de reglas a partir de colecciones de datos con una finalidad previamente determinada y para su uso en la toma de decisiones con respecto a dicha finalidad. El concepto de KDD no comprende necesariamente esta segunda parte. Esta diferencia, muchas veces inadvertida, puede ser la causa de que ambos conceptos se utilicen indistintamente en gran parte de la literatura.

Recientemente ha alcanzado gran popularidad la construcción de almacenes de datos (*Data Warehouse*, en inglés) que también puede verse traducido de otras formas, bodegón de datos, por ejemplo. Aunque un almacén es una base de datos en sí, se diferencia de esta en que contiene resúmenes, consolidaciones y análisis de la interrelación de los datos a través del tiempo. Por sus características, a un almacén de datos se accede con menos frecuencia que a las bases de datos temporales, y es la forma más simple de permitir el acceso a los datos y de facilitar la toma de decisiones sobre la base de los procesos.

Un *Data warehouse* se conforma con datos operacionales y se diseña con el propósito de facilitar la toma de decisiones. La información que se almacena en él, nunca se actualiza y sólo se habilita para consultas. Del otro lado, integra y hace consistentes a los datos extraídos de las bases de datos operacionales.

Puede resultar conveniente construir data warehouse localizados y específicos para un objetivo determinado. Estos depósitos reciben el nombre de *datamarts*.

Un enfoque que ha cobrado actualmente fuerza es el análisis en línea (en inglés, denominado *On-Line Analytical Processing u OLAP*). Se trata de una tecnología orientada al acceso y el análisis de datos en línea. Su nombre se deriva del contraste con el procesamiento de transacciones en línea (*On-Line Transaction Processing, OLTP*). Mientras que el OLTP depende de bases de datos relacionales, el OLAP ha desarrollado una tecnología de bases de datos multidimensionales. Estas bases de datos fundan los cimientos para el desarrollo de los cálculos y análisis multidimensionales que requiere la inteligencia empresarial.4

Criterios para aplicar los métodos de la minería de datos

- *Factibilidad económica - organizativa:* existe potencialmente un impacto significativo, no se conocen métodos alternativos, se dispone de personal calificado, no existen problemas de legalidad o violación de la información.
- *Factibilidad técnica:* se dispone de suficientes datos, los datos contienen rasgos relevantes, existe poco ruido en los datos y se domina la aplicación de los métodos.

Bioinformática

La bioinformática se encuentra en la intersección entre las ciencias de la vida y de la información, proporciona las herramientas y recursos necesarios para favorecer la investigación biomédica. Como campo interdisciplinario, comprende la investigación y el desarrollo de sistemas útiles para entender el flujo de información desde los genes a las estructuras moleculares, su función bioquímica, su conducta biológica y, finalmente, su influencia en las enfermedades y en la salud.5

Los estímulos principales para el desarrollo de la bioinformática son:

- El enorme volumen de datos generados por los distintos proyectos denominados genoma (humano y de otros organismos).
- Los nuevos enfoques experimentales, basados en biochips, que permiten obtener datos genéticos a gran velocidad, bien de genomas individuales (mutaciones, polimorfismos) de enfoques celulares (expresión génica).
- El desarrollo de Internet, que permite el acceso universal a las bases de datos de información biológica.

La magnitud de la información que genera las investigaciones realizadas sobre el genoma humano es tal que, probablemente, supera la generada por otras investigaciones en otras disciplinas científicas. Como se sabe, la vida es la forma más compleja de organización de la materia que se conoce. En estos momentos, los ordenadores no clasificados para uso civil más potentes del mundo (en *Celera* y en *Oak Ridge National Laboratory*, por ejemplo, con una capacidad de cálculo cercana a los 2 *Teraflops*, billones de operaciones por segundo) están dedicados a la investigación biológica, concretamente a la obtención y al análisis de las secuencias de nucleótidos de los genomas conocidos.

Ante tal situación, uno de los retos de la bioinformática es el desarrollo de métodos que permitan integrar los datos genómicos –de secuencia, de expresión, de estructura, de interacciones, etc.– para explicar el comportamiento global de la célula viva, minimizando la intervención humana. Dicha integración, sin embargo, no puede producirse sin considerar el conocimiento acumulado durante decenas de años, producto de la investigación de miles de científicos, recogido en millones de comunicaciones científicas.

La bioinformática se ocupa de la utilización y almacenamiento de grandes cantidades de información biológica, es decir, trata del uso de las computadoras para el análisis de la información biológica, entendida esta como la adquisición y consulta de datos, los análisis de correlación, la extracción y el procesamiento de la información. En otras palabras, la bioinformática es un área del espacio que representa la biología molecular computacional, que incluye la aplicación de las computadoras y de las ciencias de la información en áreas como la geonómica, el mapeo, la secuencia y determinación de las secuencias y estructuras por métodos clásicos. Las metas fundamentales de la bioinformática son la predicción de la estructura tridimensional de las proteínas a partir de su secuencia, la predicción de las funciones biológicas y biofísicas a partir de la secuencia o la estructura, así como simular el metabolismo y otros procesos biológicos basados en esas funciones. Muchos de los métodos de la computación y de las ciencias de la información sirven para estos fines, incluyendo el aprendizaje de las máquinas, las teorías de la información, la estadística, la teoría de los gráficos, los algoritmos, la inteligencia artificial, los métodos estocásticos, la simulación, la lógica, etc.

En la reunión *Chips to Hits* '99, donde hubo representantes de compañías de *software* como *Lion Biosciences*, *Informax*, *Molecular Applications Group o Gene Logic*, de empresas farmacéuticas como *Bristol-Myers* y del *National Cancer Institute* del gobierno estadounidense, se comentó que actualmente uno de los cuellos de botella de los ensayos con tecnologías basadas en *biochips* se encuentra en la carencia de herramientas bioinformáticas adecuadas para el análisis y gestión de los datos, debido a los enormes

volúmenes de datos que ellos generan. Asimismo, se resaltó la necesidad de emplear las técnicas de la minería de datos, como la mejor forma de obtener conocimientos a partir de los resultados experimentales.6

"El reto en la construcción de bases de datos es el establecimiento de una arquitectura que permita la realización de búsquedas inteligentes, la comunicación con otras bases de datos y la unión con herramientas de análisis y de minería de datos, específicas, que permitan responder a problemas biológicos concretos. Los científicos, encargados de la construcción de estas bases de datos, deben disponer de conocimientos previos que permitan determinar cuáles problemas científicos concretos necesitan una solución y cuál o cuáles métodos son los mejores para resolverlos". Así se declara en el artículo "Qué es la Bioinformática" publicado por BIOTIC. Y en la propia publicación del Instituto Carlos III de Madrid, España, se afirma: "Se necesitan herramientas para gestionar información genética en paralelo. Para ello se emplean nuevas tecnologías de extracción de conocimientos, minería de datos y visualización. Se aplican técnicas de descubrimiento de conocimientos a problemas biológicos como análisis de datos del genoma y el proteoma."7

En estos momentos, la mayoría de los proyectos que se desarrollan en el mundo en materia de genómica y proteómica, demandan la aplicación de técnicas de la minería de datos para poder determinar qué es realmente importante dentro del enorme volumen de información que se genera diariamente en el mundo. Considérese que el número total de letras (pares de bases químicas) del ADN humano ha resultado ser de 3.120 millones. El Proyecto Genoma Humano aseguró que, a los 10 años de su creación, ha terminado un primer borrador de la secuencia y completado el 85 % del ensamblaje. De los 3.120 millones de datos que componen el «libro de la vida», los científicos han encontrado que el 99,8 % son idénticos para todas las personas.

Como ha señalado *Ignacio F. Bayo*: "El principal escollo al que se enfrenta la proteómica, y en general la biología básica, es la carencia de sistemas informáticos apropiados para la inmensa cantidad de cálculos implicados en este tipo de investigaciones".8

El investigador del Consejo Superior de Investigaciones Científicas, *Alfonso Valencia*, quien se dedica al desarrollo de software para el análisis de proteínas en todos los niveles, análisis de genoma, determinación de secuencias y estructuras, así como a la comparación con bases de datos o predicción de funciones, señala: "Pese a todo, con los equipos más potentes se podría obtener mucha más información que la que se consigue, pero existe otro problema: la dispersión de los datos. Los investigadores y las empresas guardan celosamente los resultados de sus trabajos debido a la posibilidad de realizar patentes a partir de ellos. Incluso en los casos en que se coloca la información en la red es difícil trabajar con ellos porque no se han desarrollado mecanismos adecuados de búsqueda. «La base de datos de la National Library of Medicine de Estados Unidos es la mayor fuente accesible computadorizada y contiene 10 millones de referencias, pero sólo están los sumarios de los artículos técnicos, luego tiene uno que buscar lo que le interese en otros sitios...», se queja Valencia y añade: «Se trata de una información desestructurada, que no puede incorporarse directamente para estudiar la función de una proteína. Es, por así decir, una información muerta. Lo deseable sería poder cruzar datos de miles de genes o de proteínas para conseguir con rapidez indicios de su estructura y de su función que permitan avanzar en la investigación. Una posibilidad sería aplicar la tecnología que están utilizando los buscadores en el web para seleccionar cada vez con mayor precisión la información demandada, mediante análisis estadístico de las palabras claves introducidas. Ahora se empiezan a aplicar estas técnicas en el campo de la proteómica, pero aun así, resolver un proteoma, relativamente sencillo, llevará aún muchos años, decenios probablemente en el caso del proteoma humano." 8 Considérese, que si se analiza sólo desde el punto de vista cuantitativo, los componentes del DNA son cuatro nucleótidos y sin embargo, las proteínas la integran 20 aminoácidos. El aumento de volumen es evidente.

Consideraciones finales

El desarrollo de la tecnología de minería de datos está en un punto de inflexión, con respecto a su consolidación, en las aplicaciones. Existen una serie de elementos que la hacen aplicable, y una realidad que la demanda; sin embargo, existe una serie de retos que atentan contra su credibilidad. Uno de ellos es que los productos comercializados son costosos, por tanto los consumidores pueden hallar una relación costo/beneficio improductiva.

La aplicación de la minería de datos, además de permitir el descubrimiento de conocimientos para el sector comercial, soporta las investigaciones en la rama biológica, encuentran en ella una herramienta insustituible para enfrentar la avalancha de datos que producen las investigaciones genómicas y proteómicas. En este sentido, es necesario continuar elaborando herramientas computa-cionales apropiadas para su uso en varios proyectos y elevar el nivel de conocimientos sobre su utilidad para los investigadores.

Algunos de los factores que pueden crear una desilusión con las promesas de la minería de datos son:

- Que se necesite mucha experiencia para utilizar herramientas de la tecnología, o que sea fácil hallar patrones equívocos, triviales o no interesantes.
- Que no sea posible hallar patrones en tiempo o en espacio.
- Que no se establezca una adecuada comunicación en los equipos multidis-ciplinarios para elegir la herramienta adecuada y que, por lo tanto, no se alcancen los resultados esperados.
- Que existan razones organizativas, éticas o de otro carácter que impidan la utilización de toda la información necesaria para la aplicación de estas herramientas.

La primera década del siglo xxi, será un período de importancia para la aplicación de estas herramientas a gran escala.

Abstract

An extraordinary advance of the biomedical sciences will take place in the next years as a result of the Human Genome project. The new technologies based on the molecular genetics and informatics are key factors for this development, since they provide powerful tools for the obtention and analysis of genetic information. The appearance of new technologies has made possible the development of genomics, on making possible the study of the interactions of genes and their influence on the development of diseases. All this influences on the clinical diagnosis, the investigation of new drugs epidemiology and medical informatics. In the last years, data mining has experienced an increase as a support for the phylosophies of information management and knowledge, as well as for the discovery of the meaning of the data stored in big banks. This allows to explore and analyze the databases available to help in the decision-making process and it also facilitates the extraction of the information existing in the texts and to create smart systems capable of understanding them. This is commonly known as text mining. The basic components of data mining and its application to an emerging and trascendent scientifc activity, bioinformatics, are synthetically described.

Subject headings: COMPUTATIONAL BIOLOGY; MEDICAL INFORMATICS; DATABASES; DECISIÓN MAKING

Referencias bibliográficas

1. Consultoría BIOMUNDI. Estado del arte en Bioinformática. La Habana: Consultoría BIOMUNDI, 2001.
2. Morales E. Descubrimiento de conocimientos en bases de datos. [Disponible en: http://w3.mor.itesm.mx/~emorales/Cursos/KDD/node9.html]
3. Goglino D. Minería de datos. [Disponible en: http://www.infonews21.com/columnas/goglino/goglino.htm]
4. Accrue Software. An Introduction to OLAP Multidimensional Terminology and Technology. [Disponible en: http://www.accrue.com/olap/wp_intro_olap.pdf]
5. Martín Sánchez F, López Campos G, Maojo García V. Bioinformática y salud: impactos de la aplicación de las nuevas tecnologías para el tratamiento de la información genética en la investigación biomédica y la práctica clínica. Informática y Salud 1999;(19). [Disponible en: http://www.seis.es/i_s/i_s19/i_s19l.htm]
6. Parsaye K. DataMines for DataWarehouses. [Disponible en: http://www.datamining.com/]
7. Unidad de Coordinación de Informática Sanitaria (BIOTIC). ¿Qué es la Bioinformática? [Disponible en: http://biotic.isciii.es/informacion/bioinfo/definicion/queesbioinfo.htm]
8. Bayo IF. El próximo desafío se llama proteoma. [Disponible en: http://www.elpais.es/especiales/2000/genoma/descifra/proteoma.html]

Recibido: 25 de octubre del 2001
Aprobado: 13 de noviembre del 2001

Dr. *Juan Pedro Febles Rodríguez*
Centro Nacional de Bioinformática
Calle 17 No. 1420 e/n 26 y 28, El Vedado, Ciudad de La Habana. Cuba. Correo electrónico:
mailto:febles@aid.inf.cu

1 Doctor en Ciencias Técnicas. Profesor Titular en Informática Médica. Director del Centro Nacional de Bioinformática.
2 Licenciado en Bioquímica. Centro Nacional de Bioinformática.
