Minería de Datos y Genómica

Valeria Burgos

RESUMEN

El amplio acceso a computadoras de alto desempeño y dispositivos electrónicos de gran almacenamiento, entre otros, ha permitido en los últimos años la generación de cantidades masivas de datos, concepto que puede ser representado por Velocidad, Volumen y Variabilidad.

La Minería de Datos es un proceso que permite descubrir patrones o asociaciones relevantes, no plenamente descubiertas en principio con los métodos tradicionales de análisis, en grandes bases de datos y generar modelos. Para ello, usa herramientas de áreas tales como Sistemas de Bases de Datos, Almacenamiento, Aprendizaje Automático, Estadística, Visualización de la Información y Computación de Alto Desempeño.

En las últimas décadas, la biología molecular ha pasado del análisis de genes individuales a estudios más complejos que abarcan el genoma completo de un individuo. El desarrollo de tecnologías genómicas de alto desempeño, como los microarrays y la secuenciación de próxima generación (NGS), ha hecho posible producir de manera exponencial información, con la expansión de nuestro conocimiento de las bases genéticas de varias enfermedades.

En la Medicina Genómica, el uso de la Minería de Datos para el análisis de la información genómica se está convirtiendo en una necesidad cada vez más buscada, contribuyendo así hacia una medicina personalizada tal que permite inferir modelos clínicamente relevantes y definir estrategias terapéuticas individualizadas a partir de datos moleculares de pacientes.

Palabras clave: minería de datos, genómica, grandes volúmenes de datos, NGS.

DATA MINING AND GENOMICS

ABSTRACT

The availability of use of high-performance computers and large-storage electronic devices, among others, has allowed the generation of a huge masses of digital data, an idea that can be represented by velocity, volume and variety.

Data mining is a process that permits to discover relevant patterns or relations, not previously seen with traditional methods of analysis, in large databases and generate models. It uses tools from Database Systems, Data Warehouse, Machine Learning, Statistics, Information Visualization and High-Performance Computing.

In the last decades, molecular biology has moved from individual gene analysis to more complex studies that involve the complete genome. The development of high-throughput genomic technologies, such as microarrays and next-generation sequencing, has promoted the exponential growth of a huge amount of information, expanding our knowledge on the genetic basis of various diseases.

In genomics medicine, the application of data mining techniques has become an increasingly important process that contributes towards a personalized medicine, that involves the inference of clinically relevant models and defines individualized therapeutic strategies based on the molecular data of patients.

Key words: Data mining, genomics, Big Data, NGS.

Rev. Hosp. Ital. B.Aires 2016; 36(4): 160-164.

INTRODUCCIÓN

Datos. Información. Estamos en una era en la cual el volumen de los datos va creciendo de manera rápida. Los datos son hechos no organizados, crudos, que deben procesarse. Pueden ser algo simple y aparentemente sin utilidad hasta que son organizados. Cuando los datos son procesados o presentados en un contexto dado de manera tal de hacerlos útiles, se convierten en información.

Recibido 19/08/16 Aceptado 6/10/16

Laboratorio de Aprendizaje Biológico y Artificial. Departamento de Ingeniería Biomédica. Instituto de Ciencias Básicas y Medicina Experimental. Instituto Universitario Hospital Italiano de Buenos Aires.
Correspondencia: valeria.burgos@hospitalitaliano.org.ar

Diversos tipos de datos son generados por nosotros mismos en la vida cotidiana: cada vez que usamos el teléfono celular para una llamada; cuando usamos la red inalámbrica de un bar donde tomamos un café o si tildamos "Me gusta" o "etiquetamos" a alguien en las redes sociales. Cuando realizamos búsquedas en Internet. Cuando usamos la tarjeta de crédito en un comercio o efectuamos compras por Internet, si elegimos una película según demanda, o luego de realizar un chequeo médico donde los datos quedan registrados en la historia clínica electrónica.

La naturaleza de los datos generados explica la reciente explosión alrededor del concepto de *Big Data*. Definido como la colección, el almacenamiento y el análisis de cantidades masivas de datos computarizados, hoy *Big Data* crece a pasos agigantados. Datos en cantidad han existido siempre. El punto ahora es que hasta hace un

par de décadas, muchas fuentes que dan origen a esos tipos de datos no existían, así como tampoco el amplio acceso a computadoras de alto desempeño y dispositivos electrónicos de gran almacenamiento, lo cual se traduce en volumen, velocidad y variedad[1].

– Volumen: se refiere a la cantidad de datos generados. Es el tamaño (escalas de kilobytes, megabytes, terabytes, etc.) lo que determina el valor y el potencial de los datos.
– Velocidad: hace referencia a cuán rápido los datos son generados y procesados para satisfacer las demandas y desafíos del usuario.
– Variedad: se refiere a que los datos pueden ser de diferentes orígenes (textos, videos, imágenes, audio, entre otros).

En el ámbito de la salud, la cantidad de información asociada a los pacientes crece constantemente. El avance tecnológico de los últimos años ha permitido el rápido y eficaz procesamiento de enormes cantidades de muestras de pacientes, que permite evaluar la presencia de enfermedades, examinar genes o encontrar nuevos blancos para un tratamiento en particular.

Ahora bien, además del crecimiento en la capacidad computacional y de almacenamiento de los datos, los expertos indican que la verdadera revolución es que hoy se puede hacer algo con ellos. En particular, esa revolución está reflejada en el mejoramiento de métodos estadísticos y computacionales, nuevas formas de asociar bases de datos y en formas creativas de visualizar la información. De tal manera, todo esto en conjunto ayuda a encontrar el sentido a la enorme masa de información y favorece la creación del conocimiento. De esto se ocupa la Minería de Datos.

¿QUÉ ES LA MINERÍA DE DATOS?

Se conoce como Minería de Datos (*Data Mining*, en inglés) el proceso de descubrir patrones o relaciones interesantes o ambos, no plenamente descubiertos en principio con los métodos tradicionales de análisis, en grandes bases de datos. Es un campo relativamente nuevo e interdisciplinario y forma parte de un proceso aún más amplio conocido como Descubrimiento del Conocimiento (*Knowledge Discovery*, en inglés) (Fig. 1).

En general, es una secuencia iterativa formada por varios pasos:

– Preprocesamiento de los datos (detección y eliminación de valores atípicos, transformación de los datos en un formato apropiado para su análisis posterior, selección de los registros y características que son relevantes a la pregunta de interés).
– Aplicación de métodos inteligentes para extraer asociaciones o patrones interesantes (*data mining*).
– Identificación de patrones verdaderamente interesantes que representen el conocimiento adquirido.

[1]Estos tres conceptos representan "las 3 V" del desafío de analizar gran cantidad de información

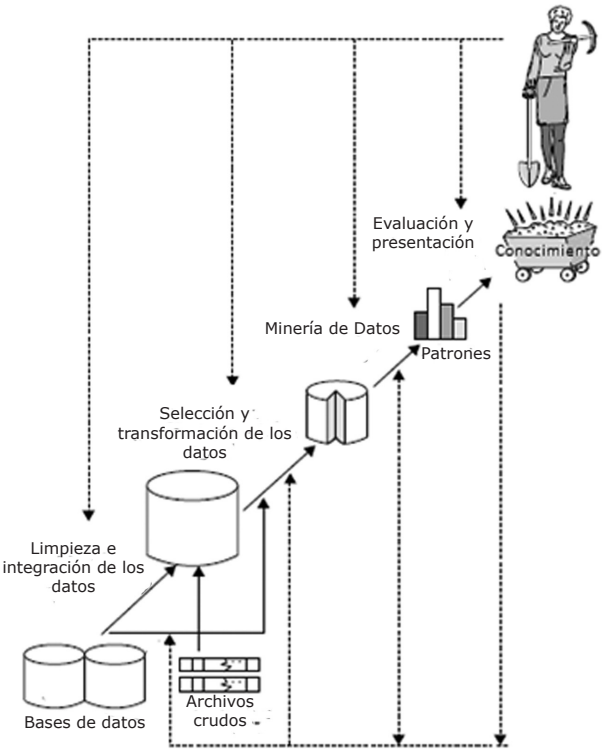

Figura 1. Minería de Datos como uno de los pasos en el proceso de Descubrimiento del Conocimiento. Adaptado de Han et al., 2012.

– Presentación del Conocimiento (aplicación de técnicas de visualización de la información de los datos analizados y presentación adecuada y comprensible al usuario).

Muy comúnmente se puede pensar la Minería de Datos como el producto del arduo trabajo de un minero, algo así como "encontrar el oro dentro de las rocas".

Como se mencionó anteriormente, la Minería de Datos es un proceso que aplica herramientas de diversas áreas tales como Sistemas de Bases de Datos, Almacenamiento, Aprendizaje Automático, Estadística, Visualización de la Información, y Computación de Alto Desempeño.

Los objetivos de la Minería de Datos pueden ser clasificados principalmente en dos categorías: descripción y predicción.

La Minería de Datos descriptiva busca descubrir información que está implícita y previamente desconocida, la cual puede ser usada en la toma de decisiones por el usuario. Para conseguir este objetivo suelen usarse técnicas de Aprendizaje Automático no Supervisadas. Este tipo de aprendizaje está caracterizado por un modelo que se ajusta a las observaciones y no hay un conocimiento previo de la naturaleza de los datos. Como ejemplos se pueden mencionar la búsqueda de patrones frecuentes en los datos, encontrar asociaciones y correlaciones interesantes entre las variables, análisis de *clusters*, análisis de valores atípicos y evolución del conjunto de datos.

El propósito de la Minería de Datos descriptiva es resumir el tipo de información que se puede obtener de los datos, limitaciones y pasos subsiguientes para seguir en el análisis. Por ejemplo, las reglas de asociación son proposiciones lógicas frecuentemente observadas en un conjunto dado de datos. Un analista puede expresar hipótesis en la forma de estas reglas y verificar su validez en los datos además de encontrar relaciones relevantes previamente desconocidas. Por otra parte, el análisis de *clusters* se refiere a formar grupos de objetos que son muy similares a otros pero diferentes de los objetos de otros *clusters*.

En los casos de predicción, la aplicación de Minería de Datos busca encontrar un modelo o función que pueda predecir alguna propiedad importante (pero no descubierta todavía) de un sistema, utilizando los datos o el comportamiento de este. En el área biomédica es de interés encontrar un modelo que prediga la aparición de síntomas antes de que aparezcan basándose en la condición de salud del paciente, su historia clínica y los tratamientos indicados. Para esto, generalmente, se pueden usar técnicas de Aprendizaje Automático Supervisadas. En este tipo de aprendizaje existe conocimiento *a priori* de los datos y, luego de un entrenamiento, se crea un modelo que puede generalizar clasificando correctamente datos pertenecientes a situaciones no vistas antes. Entre las técnicas de aprendizaje supervisado se pueden mencionar a la Regresión y la Clasificación.

El análisis de regresión es una de las técnicas estadísticas de uso más frecuente para analizar datos multifactoriales. Usa una ecuación para expresar la relación entre una variable de interés (la respuesta) y un conjunto de variables predictoras relacionadas (Montgomery et al., 2002). La salida de la función puede ser un valor numérico. En cambio, la clasificación es una técnica que predice el valor de un atributo categórico, clasifica los datos en grupos predefinidos de clases. Existen varios métodos, tales como *clustering* de Vecinos Más Cercanos, Redes Neuronales, Árboles de Decisión, entre otros.

MINERÍA DE DATOS Y GENÓMICA

El área de la biología molecular se ha visto beneficiada por un gran avance tecnológico en los últimos años. El mejoramiento de técnicas como los microarreglos de ADN[2] y la implementación de cada vez más poderosos equipos de secuenciación ha permitido generar gran cantidad de datos. En particular, la secuenciación de ácidos nucleicos es un método para determinar el orden exacto de nucleótidos presentes en una molécula de ADN o ARN. Actualmente existe un gran interés en las biociencias por los métodos de secuenciación de nueva generación (NGS, por sus siglas en inglés), los cuales secuencian de manera más rápida y menos costosa el genoma completo (Grada & Weinbrecht, 2013). También permite secuenciar profundamente partes de secuencias específicas así como analizar la diversidad microbiana en el cuerpo humano y en el ambiente. Cada equipamiento de NGS puede diferir en la técnica de secuenciación específica, gracias a lo cual se pueden abordar diversos tipos de resultados acorde con las necesidades del investigador o el servicio.

Todo esto en conjunto ha permitido avanzar en el entendimiento de ciertos fenómenos biológicos. Sin embargo, aun cuando las instrucciones están inscriptas en el genoma, no existe todavía una comprensión total sobre por qué ciertos genes se "prenden" y otros se "apagan", y cómo estos interaccionan en una gran red génica en el caso de ciertas enfermedades.

En el área de medicina genómica, el objetivo es inferir modelos clínicamente relevantes a partir de datos moleculares y dar así sustento a la toma de decisiones. Actualmente, los datos moleculares están disponibles en tres formas: 1) datos de genotipos, representados por un conjunto de polimorfismos de único nucleótido. Son alteraciones que ocurren en la secuencia genómica y que afectan a un único nucleótido. Dado que cada individuo posee muchos de esos polimorfismos en su genoma, su presencia forma un patrón único para esa persona; 2) datos de expresión de genes, los cuales pueden ser medidos por varias técnicas como microarreglos de ADN o variantes de la reacción en cadena de la polimerasa (PCR, por sus siglas en inglés). De esta manera, es posible obtener una instantánea de la actividad de los genes en un tejido en particular para un momento dado; 3) datos de expresión de proteínas, las cuales pueden ser analizadas mediante estudios a gran escala del proteoma[3] para brindar información sobre abundancia de proteínas específicas, variaciones y modificaciones. La información obtenida es útil para el armado de un perfil proteico característico en el diagnóstico, pronóstico y predicción terapéutica ante alguna enfermedad.

Una de las áreas de mayor interés en la medicina genómica es la oncología, donde hay una fuerte necesidad de definir estrategias terapéuticas individualizadas a partir de datos de expresión de genes de pacientes obtenidos mediante microarreglos de ADN. Estos consisten básicamente en superficies sólidas cubiertas por pequeños fragmentos de ADN (sondas) a las cuales se van a unir fragmentos de ADN del paciente. Luego se mide el nivel de hibridación entre las sondas y los fragmentos de interés mediante fluorescencia y, a través de un análisis de imágenes, se evalúan los niveles de expresión de los genes en la muestra. Un ejemplo que ilustra el potencial de aplicar técnicas de Minería de Datos sobre el perfil de expresión de genes al momento de predecir un tratamiento es el realizado sobre

[2]Conjunto de sondas de ADN unidas a un soporte sólido en una disposición prefijada y regular. Sirve para determinar la expresión génica de un tejido en un momento determinado, obteniendo una "foto genética transversal", según Vallin Plous, 2007.

[3]El proteoma se refiere al conjunto total de proteínas producidas por un organismo o sistema celular.

muestras de distintos tipos de cáncer de mama provenientes de una población de pacientes y un algoritmo de *clustering* jerárquico (van´t Veen et al., 2006). Mediante el uso de microarreglos que contenían 25 000 genes para estudiar, se determinó que alrededor de 5000 genes mostraron una regulación positiva en su expresión.

Usando una técnica no supervisada de *clustering* como primer acercamiento, se observó que es posible diferenciar entre tumores con "mal pronóstico" y tumores con "buen pronóstico" sobre la base del desarrollo de metástasis dentro del período de estudio. Además, cuando se relacionó este resultado con datos histopatológicos, el *clustering* no supervisado indicó la existencia de dos subgrupos de cáncer que difieren en cuanto a la infiltración linfocitaria y marcación de receptor de estrógenos. Posteriormente, los autores usaron un método supervisado de clasificación y se determinó que la firma de "mal pronóstico" está asociada al aumento en la expresión de genes del ciclo celular, invasión, metástasis y angiogénesis. El clasificador de pronóstico de tumor de mama fue validado con otro grupo de pacientes y derivó en parámetros de precisión adecuados. De esta manera, los autores destacan la importancia del clasificador desarrollado al momento de decidir el uso de una terapia hormonal adyuvante, y los genes que están sobreexpresados en los tumores con mal pronóstico son blancos para el desarrollo de nuevas drogas.

Por otra parte, dentro de la medicina genómica es de interés evaluar también parámetros no moleculares que puedan ayudar a la formación de un modelo en casos de enfermedades con bases genéticas no tan definidas. Como ejemplo, a continuación, se describe la aplicación de técnicas de aprendizaje automático supervisado en autismo.

MINERÍA DE DATOS EN AUTISMO

El Trastorno del Espectro Autista (TEA) es un síndrome definido por una amplia variedad de comportamientos, caracterizado por deficiencias en la comunicación e interacción social y comportamientos estereotipados. Los avances de investigación en el área indican un número creciente de variantes genéticas raras que parecen influir en la predisposición de un individuo con la aparición de fenotipos conductuales de TEA. Sin embargo, hasta el momento estos posibles factores de riesgo genético no están formalmente comprobados.

En este contexto, una investigación reciente ha implementado una técnica de Aprendizaje Automático para estudiar las relaciones entre los datos conductuales de pacientes diagnosticados con TEA y el genotipo (Bruining et al., 2014). Específicamente se utilizó una "máquina de soporte vectorial". En Aprendizaje Automático, las máquinas de soporte vectorial representan un conjunto de algoritmos de clasificación muy poderosos en términos de exactitud en la predicción. Para entender de manera básica la idea detrás del algoritmo de una máquina de soporte vectorial se presentan a continuación sus pasos fundamentales; en la figura 2 se indica de manera esquemática el procedimiento.

A. Un conjunto dado de elementos cuya pertenencia a una u otra clase ya está definida (cuadrado o círculo en este ejemplo) sobre la base de dos características cualesquiera (x_1 y x_2).

B. Un algoritmo de entrenamiento soporte vectorial construye un modelo que asigna nuevos ejemplos a una u otra clase mediante la separación de estos por un hiperplano[4] (indicado en línea verde).

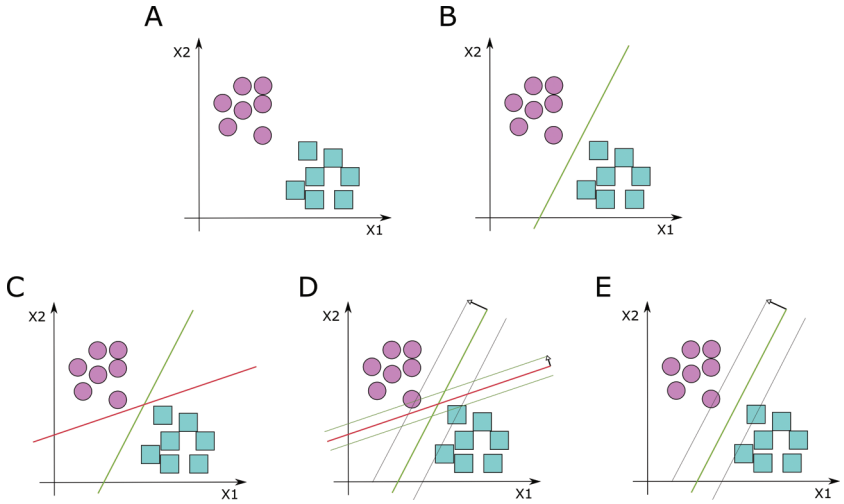

Figura 2. Representación básica de máquina de soporte vectorial.

[4]En geometría, un hiperplano es una extensión del concepto de plano. En un espacio unidimensional (como una recta), un hiperplano es un punto: divide una línea en dos líneas. En un espacio bidimensional (como el plano XY), un hiperplano es una recta: divide el plano en dos mitades. En un espacio tridimensional, un hiperplano es un plano corriente: divide el espacio en dos mitades. Este concepto también puede ser aplicado a espacios de cuatro dimensiones y más, donde estos objetos divisores se llaman simplemente hiperplanos, ya que la finalidad de esta nomenclatura es la de relacionar la geometría con el plano.

C. Sin embargo, existen varias opciones de hiperplanos (líneas verde y roja) que clasifican de manera correcta en ambas clases.
D. El objetivo de una máquina de soporte vectorial es construir un hiperplano que deje el máximo margen entre las clases (esto está indicado en líneas finas). En este ejemplo, la línea verde es la que deja mayor margen comparada con la línea roja.
E. Se puede decir que la distancia (indicada por la flecha negra) entre los elementos más cercanos y el hiperplano es maximizada cuando ciertos parámetros son definidos. Sin entrar en detalles, el margen total se puede calcular por la ecuación indicada en la figura. Minimizando este término se va a maximizar la separabilidad entre clases.

De esta manera, el modelo construido permite mapear nuevos ejemplos en el mismo espacio y, sobre la base de a qué lado del hiperplano caen, puede predecir a qué clase pertenecen.

Los autores usaron una máquina de soporte vectorial que realizó un barrido a través de una base de datos médicos de individuos diagnosticados con uno de los seis trastornos genéticos asociados al autismo: síndrome de deleción del 22q11.2 (síndrome de DiGeorge), síndrome de Down, síndrome de Prader-Willi, esclerosis tuberosa, síndrome de Klinefelter y cromosoma 15 isodicéntrico. Así, para cada síndrome, la máquina de soporte vectorial fue capaz de identificar firmas conductuales específicas y aprendió a reconocer seis trastornos genéticos asociados con esas conductas. Luego se aplicó este algoritmo de soporte vectorial a una población de pacientes con TEA sin causa conocida (autismo primario). Los resultados indicaron que el algoritmo desarrollado pudo encontrar los mismos tipos de firmas en los comportamientos de individuos con TEA primario. Es decir, una de las principales contribuciones de este trabajo es que aporta evidencia para las correlaciones genotipo-fenotipo en relación con la sintomatología autista. Aunque estos resultados son preliminares, resulta interesante que el uso de algoritmos de soporte vectorial puedan ser usados para estratificar casos de autismo primario de acuerdo con los patrones de firma conductual asociados con trastornos genéticos. De esta manera, las "firmas" de los fenotipos conductuales podrían ayudar en el entendimiento de las bases genéticas del TEA.

COMENTARIOS FINALES

Con la creciente acumulación a nivel exponencial de diversos tipos de datos biológicos, el uso de la Minería de Datos en la Genómica a gran escala se está convirtiendo en una necesidad cada vez más buscada en el área de cuidados de la salud.

Numerosos expertos en el área predicen que el futuro del cuidado de la salud está en la medicina personalizada. Este concepto introduce una nueva manera en la cual los profesionales de la salud piensan las enfermedades ya que, en vez de pensar la enfermedad en términos de sus síntomas, la medicina personalizada la considera sobre la base de las características moleculares que dirigen la enfermedad. El uso cada vez más frecuente de técnicas de Minería de Datos sobre miles y miles de datos permitirá mejorar la comprensión de esos mecanismos moleculares.

Conflictos de interés: el autor declara no tener conflicto de interés.

BIBLIOGRAFÍA RECOMENDADA

- Bruining H, Eijkemans MJ, Kas MJ, y col. Behavioral signatures related to genetic disorders in autism. Mol Autism. 2014;5(1):11.
- Grada A, Weinbrecht K. Next-generation sequencing: methodology and application. J Invest Dermatol. 2013 Aug;133(8):e11.
- Han J, Kamber M, Pei J. Data Mining. Concepts and techniques. Tercera Edición (2012), Morgan Kauffman Publishers.
- Montgomery D, Peck E, Vining G. Introducción al Análisis de Regresión Lineal. Primera edición en español (2002). Compañía Editorial Continental.
- Vallin Plous C. Microarreglos de ADN y sus aplicaciones en investigaciones biomédicas. Revista CENIC 2007;38(2):132-5.
- van 't Veer LJ, Dai H, van de Vijver MJ, y col. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415(6871):530-6.