

Minería de datos: concepto y aplicaciones

Marvin Coto-Jiménez *

Recibido: 21 de enero de 2013.

Aceptado: 15 de abril de 2013.

Resumen

Se presenta la minería de datos como un conjunto de técnicas para obtener información útil y no evidente en datos de diversos tipos. Primero se revisa el concepto y su evolución, para repasar varias de las aplicaciones actuales y describir su potencial. Especialmente en Internet. La conclusión principal es la importancia de su desarrollo y la formación de profesionales en esta área.

Palabras clave: minería de datos, bases de datos, inteligencia artificial.

1. Introducción

En el año 1889, Herman Hollerit patentó en los Estados Unidos de América una máquina tabuladora automática, que leía tarjetas perforadas. Los patrones registrados en las tarjetas distinguían los diferentes aspectos en los registros de personas. El objetivo del invento fue procesar la información del censo de 1890 en ese país, gracias a lo cual fue posible realizar esta tarea en un año en lugar de casi una década que se requería en el procesamiento manual. Esta referencia la podemos considerar como el primer desarrollo de una herramienta maquina para procesar gran cantidad de datos, lo cual constituye uno de los principios fundamentales de la minería de datos.

Como concepto, la minería de datos se utiliza de la manera que se entiende en la actualidad hasta principios de 1990. Pretende una analogía con el proceso de extraer materiales valiosos en una mina: En ésta hay gran cantidad de materiales (datos), que se procesan de forma extensa para obtener esas menores cantidades de material precioso buscado y que no era visible en el conjunto original (información útil). En el caso de la máquina de Hollerit, la minería, en su estado primitivo, consistió en llegar a conclusiones numéricas básicas, con la ventaja del tiempo que fue posible obtenerlas. La tabla 1 resume la evolución de las tecnologías utilizadas y las características del concepto desde 1960. La característica actual de ser prospectiva se refiere a su utilización para prever resultados futuros, lo cual es de sumo interés para la toma de decisiones en mercados dinámicos y de alta competencia, como los que se desarrollan en Internet. Como se describirá en la sección 3, los desarrolladores y empresas que funcionan en la red Internet tienen gran interés en los resultados que estas técnicas pueden ofrecer.

El proceso de minería de datos es visto como una evolución natural de la tecnología de la información, en el que la información se extrae de bases y almacenes de datos. La funcionalidad de los resultados buscados con ésta se pueden clasificar en dos grandes categorías [1]:

1. Funcionalidades descriptivas: Enfocadas en buscar patrones en los datos que puedan ser interpretables por el ser humano. Por ejemplo, caracterizaciones (generalidades de los datos), análisis de asociaciones (reglas presentes en el conjunto) o de agrupamientos (subconjuntos semejantes de datos) no evidentes.
2. Funcionalidades predictivas: Enfocadas en buscar predicciones basadas en inferencias, generalmente sobre modelos abstractos. Por ejemplo, clasificación y regresión (modelos que describan y distinguen subconjuntos de datos).

En la sección 3 se detallan varios ejemplos concretos de cada uno de estos tipos.

Una distinción que plantea el obtener datos adicionales, a manera de información útil, de tener solamente los datos, es el valor que éstos tienen. Contar con gran cantidad de datos, por ejemplo un banco sobre sus clientes, no garantiza que sea de utilidad para alcanzar o mejorar los objetivos de la entidad, a menos que se pueda extraer de esa información algo útil sobre lo que se pueda tomar decisiones y después evaluarlas. La cantidad de información puede ser más bien un obstáculo si no se cuenta con las técnicas y capacidad adecuada para procesarla para pasar de solo información a conocimiento en la forma de información útil.

2. Implementación y metodología

Las técnicas utilizadas en minería van desde herramientas estadísticas hasta herramientas de inteligencia artificial. De estas últimas, cuatro de las principales utilizadas son [3]:

1. Redes Neuronales: Son modelos no-lineales inspirados en las redes de neuronas. En principio consisten en un conjunto de nodos organizados en capas con determinados enlaces entre ellos. Tanto la configuración de los nodos como los valores en sus enlaces se ajustan, buscando una configuración óptima que permita obtener valores acertados en problemas de clasificación y predicción. La configuración óptima se realiza en un ajuste llamado entrenamiento, que utiliza los datos para determinar la estructura y características de

Tabla 1: Evolución de la minería de datos 1960 a la actualidad.

Evolución	Tecnologías presentes	Características
Colecciones de datos (1960-)	Computadoras, cintas, discos	Manipulación estadística.
Acceso a datos (1980-)	Bases de datos relacionales, lenguajes de búsqueda estructurados (SQL)	Resultados dinámicos de búsqueda a nivel de registros.
Almacenes de datos (1990-)	Bases de datos multidimensionales, almacenes de datos	Resultados dinámicos de búsqueda en múltiples niveles
Minería de datos (2000-)	Algoritmos avanzados, computadoras multiprocesador.	Información prospectiva y proactiva.

Fuente: Aldana, 2000 [2]

la red que mejores prestaciones dará al proceso buscado. Tienen la ventaja de que no pretenden establecer un modelo cerrado sobre un conjunto de datos, sino un modelo que pueda ajustarse y que utilice solamente los datos [4].

2. Árboles de decisión: Son estructuras en forma de árbol, con nodos organizados de forma jerárquica, y que representan conjuntos de decisiones capaces de generar reglas para la clasificación de los datos. Existen diversas técnicas para construirlos y optimizarlos, pues es crítico la eficiencia en cuanto al tiempo en que puedan realizar la clasificación, la cual depende del algoritmo utilizado y el tamaño del árbol [5]. Un ejemplo de árbol de decisión se muestra en la figura 1.

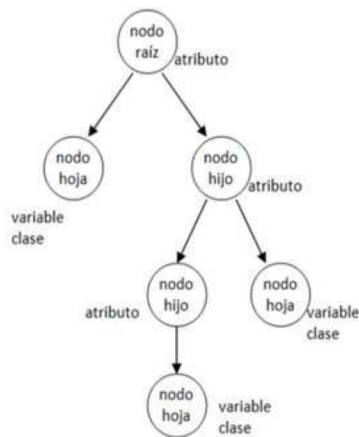


Figura 1: Ejemplo de árbol de decisión. Fuente: Barrientos *et al.* 2009.[6]

La importancia de los árboles para clasificación de datos radica en que una vez establecidos, un conjunto grande de datos puede ser procesado de forma automática y obtener resultados en tiempos relativamente cortos. En la figura 1 anterior, consis-

tirían en una clasificación -1, +1.

3. Algoritmos genéticos: Son modelos inspirados en la evolución de las especies y que se aplican generalmente en problemas de optimización. Parte de la abstracción de individuos como cadenas de números, y se establecen reglas para características como el cruce y mutación. Tienden a la optimización en la búsqueda de los mejores resultados en búsqueda y obtención de soluciones, pues es análogo a la teoría darwinista de la evolución en el sentido de la supervivencia de los individuos más aptos (mejores soluciones) como preferencia a los menos aptos [7].
4. Vecinos más cercanos: Se denomina así a la técnica de agrupación de datos que permite clasificarlos de acuerdo con su similitud, utilizando una medida de cercanía que puede partir de la representación de los puntos en una gráfica [8]. Existen diferentes maneras de realizarlo, y de forma semejante a las anteriores, la intención principal es establecerlo como algoritmo que pueda ser programado en computadora, para poder automatizarlo. En la figura 2 se muestra un ejemplo de agrupamiento por vecinos más cercanos, para el caso de tres grupos claramente diferenciados en su representación gráfica en dos dimensiones, pero que puede no ser evidente a partir del conjunto de datos original. Es posible que el conjunto original tenga dimensión mayor a dos y que haya sido necesario aplicar métodos de reducción de dimensión, lo cual agregaría complejidad al establecimiento de los grupos.
5. Reglas de inducción: Se trata de extraer reglas de la forma *si-entonces* de un conjunto de datos, combinadas e incluso utilizando variables negadas. [9]

En principio, estas técnicas no son específicas para un tipo de datos, y tienen aplicaciones diversas

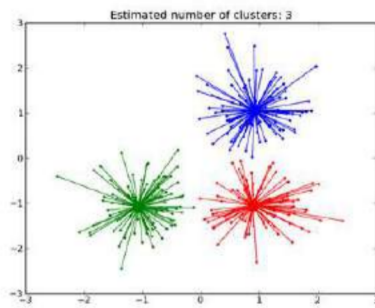


Figura 2: Ejemplo de agrupamiento. Fuente: Brendan 2007 [10].

fuera del tema de minería, por lo que usualmente se prueban varias en varias configuraciones y algoritmos para determinar las mejores. No se cuenta con un patrón o regla exacta sobre la aplicabilidad óptima de cada una, por lo que los algoritmos pueden ser diferentes para cada tipo y cada caso.

En este tema, las técnicas son parte de la minería, una herramienta de ésta. Su solo uso no identifica un proceso de minería de datos, pues pueden tener objetivos diversos en otras áreas.

Por otra parte, es importante considerar cuáles tipos de datos y con qué características pueden ser sujetos de las técnicas de minería de datos. Algunos de los tipos se describen a continuación [11]:

- Archivos de texto: En la actualidad el tipo de datos más común, consiste en texto simple o en formato binario, con datos que provienen de transacciones bancarias, series de tiempo, resultados científicos experimentales, entre otros.
- Bases de datos relacionales: Son tablas que contienen entidades o atributos de estas entidades. Consisten en columnas y filas, que organizan atributos y tuplas, respectivamente.
- Bases de datos transaccionales: Es un conjunto de datos que representan transacciones, cada una con una estampa de tiempo que registra el momento de la transacción, un identificador y una serie de ítems. Puede tener asociadas descripciones.
- Multimedia: Incluyen vídeo, imágenes y audio. Se pueden almacenar en bases de datos relacionales extendidas u orientadas a objetos. En general se caracteriza como datos de alta dimensionalidad.
- Bases de datos especiales: Además de información en alguna de las formas anteriores, contie-

nen información geográfica tal como mapas o registros de posicionamiento global.

- Series de tiempo: Contienen datos que provienen de un flujo continuo de información, tales como datos de mercados o actividades en sitios web que evolucionan continuamente con el tiempo.
- Internet: Es el repositorio de datos más heterogéneo y dinámico. En general la información está organizada en documentos interconectados, de diferentes tipos, por ejemplo texto, audio. Imágenes. Se puede considerar que esta información está compuesta de cuatro niveles: Los contenidos, la estructura, la relación entre ellos y la dinámica o evolución de éstos. En la actualidad es una de las fuentes de datos de mayor interés, tanto por su propia complejidad, como por las ventajas que puede representar la extracción de información útil de la gran cantidad de datos, que puede ser aprovechada en estudios de mercado, recomendaciones a usuarios, predicciones, entre otros. Se vuelve entonces importante desde el punto de vista social y económico [12].

En cuanto a método, la minería de datos es un proceso que invierte la dinámica del método científico [13], pues en éste, primero se formula una hipótesis y se diseña un experimento para comprobarla o rechazarla. En minería de datos, la hipótesis viene después del análisis de los datos, la cual puede ser comprobada con los datos mismos, lo que no resultaría una comprobación experimentalmente válida. Esta consideración es importante para enmarcar los resultados esperados de la minería y diferenciarlos de un conocimiento científico estándar.

Además de identificar los tipos de datos sujetos a la minería, es de valor contar un proceso estructurado para aplicarla, de valor si se piensa en automatizar su proceso, sobre todo en medios muy dinámicos donde se requiere la toma de decisiones y acciones a una velocidad cercana a lo que se conoce como tiempo real.

El mencionado proceso, por etapas, se puede sistematizar en cuatro fases [13]:

1. Filtrado de datos: Se refiere a una distinción y selección previa, pues pocas veces es posible aplicar un algoritmo sobre los datos tal como están registrados. Lo más usual es que se requieran corregir, completar y eliminar los no válidos.
2. Selección de variables: Dado que la cantidad de datos usualmente es muy grande, después del filtrado se encuentra ante gran cantidad de regis-

tros. Seleccionar variables tiene como objetivo elegir las variables más importantes en el caso para reducir su tamaño. Para este se eligen los mejores atributos del problema o se utilizan algoritmos heurísticos, de distancia o probabilidad para buscar variables independientes.

3. Algoritmos de extracción de conocimiento: Mediante las técnicas específicas (como las técnicas de inteligencia artificial citadas al principio de la sección), ya sean individuales o combinadas, se obtiene un modelo que representa patrones en las variables.
4. Interpretación y evaluación: Cuando se obtiene el modelo, se somete a un proceso de validación.

La intención del proceso es que los resultados puedan validarse y representen conclusiones no evidentes de los datos. Usualmente estas conclusiones no son detectables en los datos por la gran cantidad de éstos. La parte de interpretación y evaluación puede realizarse sin supervisión de seres humanos, por ejemplo, en el contexto de Internet si se requiere llegar a conclusiones inmediatas sobre las acciones que un usuario está realizando y las características que representa.

3. Aplicaciones

Una vez analizadas las fases requeridas y los tipos de datos sobre los que se puede actuar, se procede a revisar qué problemas ha resuelto la minería de datos, tales que la hacen de interés creciente sobre todo en el plano de investigación empresarial y en Internet. Se listan, como ejemplo, los siguientes [14]:

- Generación de recomendaciones: Sugerir un producto después de que el cliente seleccionó uno o más. El proceso lo requiere debido a la gran cantidad de clientes que pueden existir, y a la no evidente relación que existe en la preferencia de productos, la cual se puede basar en semejanzas con otros clientes que han realizado más de una compra.
- Detección de anomalías: La minería de datos puede detectar un dato anómalo, que puede corresponder a datos no deseados o incluso peligrosos, como fraudes en cuentas bancarias.
- Análisis de "separación": En operaciones financieras y bancarias, se refiere a la detección de un cliente que muestra un comportamiento temprano de que puede ser un signo de tendencia a cambiar sus preferencias o relaciones comerciales con las entidades actuales.
- Administración de riesgos: Determinar riesgos en

inversiones y créditos es de vital importancia, pues indicadores económicos y ventajas competitivas dependen de un cálculo adecuado de los riesgos en las operaciones.

- Segmentación de clientes: El conocimiento de las características de los clientes y el establecimiento de grupos que compartan estas características permite administrar campañas direccionadas a ciertos grupos que les permita situar mejor sus productos.
- Publicidad dirigida: De especial interés en Internet, donde es posible registrar las actividades de los clientes en ciertas páginas (como Amazon o Facebook), establecer patrones de comportamiento y posibles intereses, de manera que se pueda ofrecer a los anunciantes garantía que sus productos estarán a la vista de clientes o compradores potenciales.

En Internet, además, se rescata el término Minería en Web para especificar el interés específico de aplicar estas técnicas "para el descubrimiento y extracción automática de informaciones de documentos y servicios de la World Wide Web" [15]. Específicamente en aplicaciones como:

- Identificar páginas Web de mayor interés.
- Clasificar documentos.
- Clasificar los resultados de los motores de búsqueda.
- Identificar y recuperar eficientemente documentos.
- Encontrar regularidades en los registros de actividad.
- Personalizar las informaciones que se muestran.
- Identificar los intereses, preferencias e intenciones de usuarios.
- Sugerir términos de búsqueda.
- Realizar recomendaciones de acuerdo con perfiles de usuarios.
- Asistir a investigadores en búsquedas.

Como se mencionó, la cantidad y diversidad de información son un reto para las técnicas utilizadas, pues puede ser necesario otras técnicas como visión por computador, interpretación de imágenes y procesamiento de lenguaje natural.

4. Minería en grandes cantidades de datos (big data)

El concepto de Big Data se refiere a cantidades enormes de información que se obtienen, tanto generada por seres humanos o por computadoras, como registros de actividad, transacciones electrónicas, búsqueda-

das de Internet, datos de correo electrónico, y actividades en redes sociales. En el año 2010, por ejemplo, se crearon 1.5×10^{21} bytes de estos datos en Internet [16]. Debido a lo dinámico de la información y la velocidad con que se genera, el interés es poder encontrar tendencias, relaciones y en general hipótesis válidas que permitan tomar decisiones, estrategias de desarrollo y de mercado para obtener ventajas.

Para las empresas, el aprovechamiento de la minería para datos de estas características representa retos tanto en infraestructura de almacenamiento y procesamiento como en la captación de personal capacitado que pueda adaptar e innovar para las aplicaciones específicas. Solamente en Estados Unidos de América, se estima que en el año 2018 habrá una necesidad de 140 000 a 190 000 expertos con estos conocimientos [17]. Los resultados de la minería en Big Data pueden tener un impacto tan grande como determinar relaciones de causalidad entre eventos económicos y noticias generadas en múltiples campos [18].

Se trata entonces de una nueva tendencia con grandes posibilidades de desarrollo e importantes retos por su complejidad y la velocidad con que requiere los resultados.

Conclusiones

La minería de datos es un conjunto de técnicas para extraer información útil de grandes cantidades de datos en cortos periodos de tiempo. Ha sido de especial importancia desde la aparición de las computadoras, y ha evolucionado de forma paralela a éstas, gracias a la facilidad de realizar cálculos cada vez más rápido, y la creciente cantidad de información generada en redes de computadoras y entidades financieras, empresariales y comerciales.

Se destaca la diversidad de aplicaciones donde ha mostrado su utilidad y el interés cada vez mayor por implementarla y aprovechar las ventajas que ha mostrado, lo cual la puede colocar entre las áreas prioritarias para la formación de profesionales en el campo de tecnologías de la información.

Su uso actual y potencial en Internet plantea retos de interés para la investigación, con resultados que podrían verse a corto plazo y aportar al cambio en la utilidad de las redes para fines académicos y comerciales.

Referencias

1. Daylan, A.; Bertan, Y; y Badur, O. An Experimental Study for Extending Data Mining Standards.

- International Review on Computers and Software (I.R.E.CO.S.)*, Vol. 6(3): 336-346 (2011).
2. Aldana, W. Data mining industry: Emerging New Trends and New Opportunities. *Master Engineering in Electrical Engineering and Computer Science at the Massachusetts Institute of Technology*, 2000.
3. Garrido, L. y Latorre, J. Aplicaciones empresariales de Data Mining. *QUESTIO*, Vol. 25(3): 499-508 (2001).
4. Nirkhi, S. Potential use of artificial neural network in data mining. *The 2nd International Conference on Computer and Automation Engineering (ICCAE)*, IEEE. Vol. 2: 339-343 (2010)
5. Grünig, R. y Kühn, R. *Successful Decision-making. A Systematic Approach to Complex Problems*. Segunda edición, Springer 2009.
6. Barrientos, Rocio *et al.* Árboles de decisión como herramienta en el diagnóstico médico. *Revista Médica UV* Vol 9(2): 19-24 (2009).
7. Coello, C. *Introducción a la Computación Evolutiva*. CINVESTAV-IPN. 2012. Disponible en: <http://delta.cs.cinvestav.mx/~ccoello/genetic.html>.
8. Gong, A. y Lui, Y. Improved KNN Classification Algorithm by Dynamic Obtaining K. *Advanced Research on Electronic Commerce, Web Application, and Communicatio. International Conference, EC-WAC* (2011).
9. Grzymala-Busse, J.W.: Rule induction. *Data Mining and Knowledge Discovery Handbook*, 2nd edn., pp. 249-265. Springer, Heidelberg (2010).
10. Brendan J. y Delbert D. Clustering by Passing Messages Between Data Points, *Science*. Feb. 2007.
11. Zaïane, O. *Principles of Knowledge Discovery in Databases*. University of Alberta. Department of Computing Science. 1999.
12. Baeza-Yates, R. Tendencias en minería de datos de la Web. *El profesional de la información*. Vol 18(1): 5-10 (2009)
13. Vallejos, S. *Minería de Datos*. Universidad Nacional del Nordeste. Argentina 2006.
14. Tabladillo, M. ¿Por qué utilizar minería de datos? *The SolidQTM Journal*, Julio 2010-www.solidq.com/sqj
15. Pagola, J. *Estado del Arte del Web*. Centro de Aplicaciones de Tecnologías de Acanzada, Cuba. 2007.
16. Merian, L. IT looks for new tools to exploit 'Big Data'. *Computer World*, 2011.
17. Stackpole, B. Your Big Data To-Do List. *Computer World*, Feb. 13 2012.
18. Lamont, J. Big data has big implications for knowledge management. *KmWorld*, April 2012.