Data mining

Enric Mor i Pera (coordinador)
Ramon Sangüesa i Solé
Luis Carlos Molina Félix

PID_00159225

Material docente de la UOC

Universitat Oberta de Catalunya
www.uoc.edu

Enric Mor i Pera

Ingeniero en informática por la Universitat Politècnica de Catalunya. Profesor de los Estudios de Informática y Multimedia de la UOC. Está realizando la tesis doctoral en la Universitat Oberta de Catalunya, donde ha obtenido el Diploma de estudios avanzados. Las áreas de investigación incluyen la interacción persona ordenador, *e-learning y data mining*.

Luis Carlos Molina Félix

Ingeniero en Electrónica por el Instituto Politécnico Nacional (México). Máster en Ciencias de la Computación y Estadística, Universidad de Sao Paulo (Brasil). Actualmente cursa el doctorado en Inteligencia Artificial en la UPC.

Ramon Sangüesa i Solé

Doctor en Informática por la Universidad Politécnica de Cataluña. Especializado en Inteligencia Artificial, Aprendizaje Automático y Agentes Autónomos sobre Internet. Profesor titular en la Facultad de Informática de la UPC.

Primera edición: septiembre 2010

Av. Tibidabo, 39-43, 08035 Barcelona
Diseño: Manel Andreu
Realización editorial: Eureca Media, SL
Depósito legal: B-28.199-2010
ISBN: 978-84-693-4219-0

Introducción

En los últimos años ha habido un incremento sin precedentes en la informatización de las organizaciones y en sus capacidades de transmisión de datos en todos los ámbitos de actividad.

Ejemplo del alcance del incremento en la capacidad de cómputo

Los datos generados por una sola de las últimas misiones de la NASA equivale en número de bytes a 100 veces las de todas las misiones de la NASA hasta el momento; y estos datos se tienen que analizar. En concreto, el sistema de satélites *Earth Orbiting Systeme* (EOS) de la NASA puede generar hasta 50 Gigabytes de datos cada día (Way, 1991).

Al mismo tiempo, se ha reconocido como importante para la ventaja competitiva de las organizaciones su capacidad de gestionar y crear conocimiento, basándose en la información generada dentro de la organización y captada de su entorno. Se disponen de más y más datos respecto del propio funcionamiento de cada empresa, sobre las transacciones con sus clientes y proveedores y las estadísticas sobre otros fenómenos de interés (por ejemplo, las bases de datos metereológicas sobre condiciones climáticas, sequías, etc., obtenidas por satélite son interesantes para quien negocia con productos agrícolas). Para ampliar información sobre este punto podéis consultar Fayyad, 1996.

La base de datos de la Mobil Oil

La Mobil Oil ha recogido y continúa recogiendo datos en una gigantesca base de datos sobre exploración de petróleo con el fin de detectar nuevos lugares de extracción, mejorar la eficiencia de los métodos, etc. El volumen de esta base de datos se calcula alrededor de los 100 TB en la actualidad y continúa creciendo (Brachman, 1996).

Data mining en la gestión de empresas

Cadenas como Wal-Mart (USA), Sainsbury's (Gran Bretaña) o Carrefour (Francia) gestionan alrededor de 20 millones de transacciones al día que se almacenan y se analizan (Babcock, 1994).

Las grandes empresas de distribución comercial, en particular las compañías que operan sobre grandes superficies comerciales, están interesadas en conocer y anticipar la evolución de las ventas y pautas de comportamiento de sus clientes, sus diferencias regionales, etc., con el fin de poder anticipar las necesidades, programar ofertas y planificar las compras (Anand, 1998; Berry, 1997). Utilizan una gran cantidad de datos procedentes de los movimientos de materiales, así como del análisis de cada una de las compras de cada uno de los clientes (y analizan los datos recogidos en las cajas o terminales de punto de venta). Procesos parecidos de recogida de datos sirven para llevar a cabo la gestión financiera (Babcock, 1994) o tomar decisiones con respecto a la concesión de privilegios para los clientes (tarjetas de crédito, por ejemplo; consultad Carter, 1987).

Esta manera de reconocer el valor que pueden tener los datos se extiende también a otros ámbitos. Por ejemplo, los grandes sistemas hospitalarios recogen y analizan datos con el fin de minimizar los costes debidos a la prolongación innecesaria de la estancia de los pacientes, o bien localizan problemas en el

tratamiento, o relaciones entre complicaciones postoperatorias y un determinado tipo de intervenciones.

Este mismo enfoque se utiliza para establecer los patrones de acceso de los visitantes de distintas webs (Viveros, 1997; Sangüesa, 1997), conocer qué temas les interesan y ofrecerles recomendaciones. Ejemplos sencillos de este tipo de funcionamiento se pueden encontrar en varios negocios virtuales, como la librería Amazon. Con el aumento del comercio electrónico, disponer de datos con respecto a los patrones de conducta de los clientes virtuales es cada vez más importante (Bussiness Week, 1999), más aún si consideramos, por ejemplo, la aparente facilidad con que se pueden observar las pautas de navegación a partir del análisis de los ficheros de acceso y grabación de datos (ficheros de *log*).

Data mining en la NBA

Como anécdota, pero muy reveladora, los entrenadores de la liga profesional americana de baloncesto utilizan un producto de *data mining*, InfoScout (IBM), para analizar todos los partidos de una temporada y poder decidir en un momento dado qué configuración de jugadores puede darles ventaja en una situación de juego concreta. Los resultados aseguran de dos a cuatro puntos más por partido, lo que, en una liga como ésta, puede tener todo tipo de consecuencias.

Existe, pues, una extendida lógica de apreciación del valor que los datos pueden tener. Consiguientemente, se han generado nuevas necesidades en las empresas, que se reflejan en la creación de nuevos perfiles profesionales correspondientes a las tareas de *data mining*. Y esta tendencia continúa y se amplía. La racionalidad que actúa detrás de este proceso se puede aplicar perfectamente a ámbitos más modestos. De hecho, las pequeñas empresas que adoptan esta manera de considerar los conjuntos de datos, quizás no tan grandes como los que hemos apuntado anteriormente, adquieren rápidamente una nueva forma de ver y organizar sus propios procesos de trabajo para mejorarlos.

Es importante darse cuenta del hecho de que toda esta acumulación y procesamiento de datos está dirigida para obtener conocimientos de importancia estratégica o táctica para las diferentes organizaciones (Arnand, 1998). Por ejemplo, si una gran superficie conoce mejor las pautas de compras de temporada y geográficas de sus clientes, puede mejorar el servicio y los beneficios. Sin aprovechar el conocimiento que está escondido bajo muchos datos no lo podría conseguir. Encontrar este "conocimiento oculto" o no hacerlo tiene importantes repercusiones económicas y organizativas. Aplicarlo o no, puede marcar diferencias definitivas respecto a la competencia y representar ganancias o pérdidas millonarias. Por este motivo, este descubrimiento se asimila a la tarea de la minería, que tiene que retirar toneladas de escombros (los datos) para encontrar el mineral (el conocimiento).

La paradoja reside en el hecho de que aunque hay más capacidad para generar y tratar de forma automática y rutinaria los datos, es decir, de estar en

condiciones óptimas para obtener buenos análisis, el volumen mismo de los datos dificulta su tratamiento rápido y eficiente para la toma de decisión. Por este motivo, es muy importante el desarrollo de nuevas herramientas que aceleren el proceso de análisis y que lo adapten tan dinámicamente como sea posible a las necesidades cambiantes de quienes tienen que tomar decisiones: un reto fenomenal para todos los que trabajan en estadística, algorítmica, inteligencia artificial, aprendizaje automático, bases de datos y sistemas de información.

Para alcanzar un buen nivel de conocimientos en esta nueva actividad de *data mining*, se debe ser capaz de combinar conceptos y métodos procedentes de diferentes disciplinas. Por ejemplo:

- De las bases de datos (que se dedican a mejorar la captación, la organización y la consulta de datos, pero no se habían propuesto de buen principio la disposición correcta para la extracción del conocimiento).

- De la estadística, que tiene una probada tradición en la extracción de modelos a partir de datos, pero que no se había planteado la problemática de extraerlos a partir de bases de datos con una organización complicada, o de interactuar de forma continua con el usuario o de extraer modelos más simbólicos que numéricos.

- Del aprendizaje automático, una subdisciplina de la inteligencia artificial que se había encargado de extraer conocimiento de observaciones y ejemplos, pero que había simplificado las fuentes de datos, obviando las oportunidades de aprendizaje implícitas en las complejidades propias de las bases de datos reales.

En conjunto, *data mining* es una empresa que permite aprender muchas cosas de diferentes campos e integrarlas bajo una misma perspectiva al servicio de los objetivos prácticos concretos.

Esta asignatura intenta dar a conocer las diferentes aplicaciones de *data mining*, los métodos que se pueden utilizar en cada caso y, sobre todo, desarrollar una actitud, que podríamos llamar *de disposición alerta o atenta*, para detectar cuándo es necesario emprender un proyecto de *data mining* y cómo organizarlo.

Es crucial darse cuenta de que *data mining* es más una metodología y un proceso continuo que una técnica o un conjunto de técnicas. Por lo tanto, es muy importante reforzar la mencionada actitud exprimiendo el caso práctico que actúa como hilo conductor a lo largo de los diferentes módulos y los otros casos de estudio que se presentan en la asignatura de tal forma que se pueda interiorizar el proceso de forma adecuada.

Objetivos

Con los materiales didácticos asociados a esta asignatura el estudiante alcanzará los objetivos siguientes:

1. La internalización del proceso de *data mining*.

2. El refuerzo del proceso para la práctica con casos "reales".

3. El conocimiento de cada una de las técnicas que sirven al objetivo final de los distintos proyectos de *data mining* posibles.

4. El conocimiento de dos fases normalmente despreciadas en asignaturas de este tipo como la preparación de los datos y la evaluación de los modelos finales obtenidos.

Contenidos

4. Métodos de agregación
5. Interpretación de los modelos obtenidos
6. Ponderación de los métodos de agregación

Módulo didáctico 6

Reglas de asociación

Ramon Sangüesa i Solé; Luis Carlos Molina Félix

1. ¿Qué son las reglas de asociación?
2. Ponderación de las reglas de asociación

Módulo didáctico 7

Redes bayesianas

Ramon Sangüesa i Solé

1. ¿Qué son las redes bayesianas?
2. Métodos de construcción de redes bayesianas a partir de datos
3. Clasificación con redes bayesianas

Módulo didáctico 8

Evaluación de modelos

Ramon Sangüesa i Solé

1. Evaluación de modelos

Módulo didáctico 9

Caso de estudio: pozos de petróleo

Luis Carlos Molina Félix; Ramon Sangüesa i Solé

1. El problema
2. Metodología
3. Resultados

Bibliografía

Bibliografía básica

Adriaans, P.; Zantinge, D. (1996). *Data Mining*. Addison-Wesely Longman Limited.
Libro muy sencillo y claro, aunque superficial. Suficiente para tener una primera idea de los objetivos, la problemática, métodos y técnicas de la disciplina. Comenta muy brevemente algunos casos prácticos muy interesantes.

Berry, M.J.A.; Linofff, G. (1997). *Data Mining Techniques for Marketing, Sales and Customer Support*. Nueva York: Wiley Computer Publishing, John Wiley & Sons.
Superficial, pero adecuado en la descripción de técnicas dirigidas a *marketing* y relación con clientes.

Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (ed.) (1996). *Advances in Knowledge Discovery and Data Mining*. AAAI Press/MIT Press.
Más académico que Adriaans (1996), este libro recopila artículos que exponen las técnicas más conocidas y discute detenidamente varios aspectos teóricos. Incluye descripciones de los aspectos prácticos de implementación de sistemas reales.

Weiss, S.M.; Indurkhya, N. (1998) *Predictive Data Mining: A practical guide*. San Francisco: Morgan Kaufmann.
Muy dirigido a los problemas de predicción, pero muy bien tratado aunque en ocasiones de forma demasiado sucinta. Se completa con un CD con programas y conjuntos de datos para practicar.

Witten, I.H.; Frank, E. (2000). *Data Mining. Practical Machine Learning Tools and Techniques with JAVA implementations*. Morgan Kaufmann.
Excelente. Combina muy adecuadamente la presentación de las técnicas procedentes de aprendizaje automático con las que proceden de la estadística y trata de forma adecuada la preparación de datos y evaluación de modelos. Además, incluye librerías de JAVA para implementar los métodos descritos.

Bibliografía complementaria

ACM (1996). "Special Issue on Data Mining". *Communications of ACM* (noviembre).

Anand, S.; Büchner, A. (1998). "Decision Support Using Data Mining". *Financial Times Pitman Publishing*.

Babcock, C. (1994). "Parallel Processing Mines Retail Data". *ComputerWorld*, (núm. 6, 26 de septiembre de 1994).

Barr, D.I.; Mani, G. (1994). "Using Neural Nets fo Manage Investments", *AI Expert* (febrero, pág. 16-21).

Bertold, M.; Hand, D. (1999). *Intelligent Data Analysis: An Introduction*, Springer Verlag.

Bhandari, I.; Colet,E.; Parker, J.; Pines, Z.; Pratap, R.; Ramanujan, K. (1997). *Advanced Scout: Data Mining and Knowledge Discovery in NBA Data*. Data Mining and Knowledge Discovery (vol. 1, núm. 1, pág. 121-125).

Brachman, R.; Khabaza, T.; Kloesgen, W.; Piatetsky-Shapiro, G.; Simoudis, E. (1996). "Industrial Applications of Data Mining and Knowledge Discovery". *Communications of ACM* (noviembre).

Business Week (1999). *The Information Gold Mine* (26 de julio).

Carter, C.; Catlett, J. (1987). "Assessing Credit Card Applications Using Machine Learning". *IEEE Expert* (pág. 71-79, otoño).

Cios K.; Pedrycz, W.; Swiniarski, R. (1998). *Data Mining Methods for Knowledge Discovery*. Kluwer.

Elder, J.F. (IV); Abbott, D.W. (1998). *A Comparison of Leading Data Mining Tools*. Tutorial desarrollado en la conferencia KDD-98. Nueva York.

Fayyad,U.; Haussler, D; Sotolorz, P. "Mining Science Data". *Communications of ACM* (noviembre).

IEEE (1996). *IEEE Transactions Knowledge and Data Engineering* 8(6), diciembre 1996, *Special Section On Mining Of Databases*.

IEEE (1999). Bramer M.A. (ed.). *Knowledge Discovery and Data Mining: Theory and Practice*. IEEE Books.

IEEE (1999). "Special Issue on Data Mining". *IEEE Computer* (vol. 32, núm. 8, agosto).

Kohavi, R.; Provost, F. (1998). "Special Issue on Applications of Machine Learning and the Knowledge Discovery Process". *Machine Learning Journal* (abril).

Mena, J. (1999). *Data Mining Your Website*. Digital Press.

Sangüesa, R.; Saxon, S.; Nicolás, M.; Cortés, U. (1997). "WebProfile or Agents the other Way Round". *Proceedings of the First International Workshop on Applications of Agents to Web-based Information Systems* (noviembre). México.

Viveros, M.S.; Elo-Dean, S.; Wright; Duri, S.S. (1997). "Visitor's Behaviour: Mining Web Servers". *Proceedings of the First International Conference on the Practical Application of Knowledge Discovery and Data Mining, PADD-97*. Londres.

Way, J.; Smith, E.A. (1991). "The Evolution of Synthetic Aperture Radar Systems and their Progression to the EOS SAR". *Transactions on Geoscience and Remote Sensing* (núm. 29, vol. 6, pág. 962-985). IEEE.

Fayyad, U.M.; Mannila, H.; Piatetsky-Shapiro, G. (ed.). *Data Mining and Knowledge Discovery*. ISSN: 1384-5810. http://www.kluweronline.com/issn/1384-5810.

Quizá la más amplia y que reúne más contenido (artículos, presentaciones en transparencias y *software* de uso público) y referencias a otras páginas sea:

kdnuggests. *http://www.kdnuggets.com* y su lista de distribución, mantenida por Gregory Piatetsky-Sahpiro (subscripción: *kdd-request@gte.com*).

SIGKDD Explorations. Boletín del *Special Interest Group on Knowlege Discovery from DataBases* (ACM). *http://www.acm.org/sigkdd*.

Neuronet. *http://www.k.cl.ac.uk/neuronet*.
Información sobre paquetes de dominio público para construir redes neuronales.