Data Mining & Big Data

Tecnologías tradicionales de almacenamiento de datos

- *Guillermo Bonafonte Criado*

Universidad Pontificia Salamanca

Índice

Introducción

El almacenamiento de datos se ha convertido en una tarea rutinaria de los sistemas de información de las organizaciones. Esto es aún más evidente en las empresas de la nueva economía, el e-comercio, la telefonía, el marketing directo, etc. Los datos almacenados son un tesoro para las organizaciones, es donde se guardan las interacciones pasadas con los clientes, la contabilidad de sus procesos internos, representan la memoria de la organización. Pero con tener memoria no es suficiente, hay que pasar a la acción inteligente sobre los datos para extraer la información que almacenan. Este es el objetivo de la minería de datos.

¿Qué es?

La minería de datos tiene como propósito la identificación de un conocimiento obtenido a partir de las bases de datos que aporten hacia la toma de decisión.

Minería de datos es la exploración y análisis de grandes cantidades de datos con el objeto de encontrar patrones y reglas significativas que aporten conocimiento.

Es un mecanismo de explotación que consiste en la búsqueda de información valiosa en grandes volúmenes de datos.

• Ligada a las bodegas de datos (warehouse) con la cual los algoritmos de minería de datos obtienen información necesaria para la toma de decisiones.

Estadística VS Data Mining

	Estadistica	**Mineria de datos**
Construcción de modelos	Ceñido a premisas y teoremas	Mayor libertad en la construcción, interpretable
Búsqueda	Test de la razón de la verosimilitud	Metaheurísticos
Transparencia	Más complicados de interpretar	Más claros y sencillos
Validación	No	Sí

¿Qué problemas aborda la minería de datos?

Cualquier problema para el que existan datos históricos almacenados es un problema susceptible de ser tratado mediante técnicas de Data Mining.

Búsqueda de asociaciones

Un cierto suceso, ¿está asociado a otro suceso?, ¿podemos inferir que determinados sucesos ocurren simultáneamente más de lo que seria esperable si fuesen independientes?, ¿es posible sugerir un producto, sabiendo que otro ha sido adquirido?

Detección de ciclos temporales

Todo consumidor sigue un ciclo de necesidades que ocasionan actos de compra distintos a lo largo de su vida. Detectar los diferentes ciclos y la fase donde se situ´a cada consumidor ayudara´ a crear complicidades y adecuar la oferta de productos a las necesidades y crear fidelización.

Predicción

A menudo deberemos efectuar predicciones: ¿cuál es la probabilidad de baja de un cliente?, ¿cuál es el precio de una vivienda concreta?, ¿lloverá mañana? Es- tas y muchas más son preguntas que deberemos responder, para ello construiremos un modelo a partir de los datos históricos. Si la variable de respuesta es continua (p. e. la rentabilidad de un cliente) diremos que se trata de un problema de regresión, mientras que si la variable de respuesta es categórica (p. e. la compra o no de un producto) diremos que se trata de un problema de clasificación.

Técnicas

En general, cualquiera que sea el problema a resolver, no existe una única técnica para solucionarlo, sino que puede ser abordado siguiendo aproximaciones distintas. El número de técnicas es muy grande y sólo puede crecer en el futuro. También aquí, sin pretender ser exhaustivos, la siguiente es una lista de técnicas con una breve reseña.

Supervisados

Predicen el valor de un atributo de un conjunto de datos conocidos otros atributos.

- Clasificación, Predicción
- Ejemplos: Algoritmos genéticos, redes bayesianas, redes neuronales, árboles de decisión, regresión.

Redes bayesianas

Consiste en representar todos los posibles sucesos en que estamos interesados mediante un grafo de probabilidades condicionales de transición entre sucesos. Puede codificarse a partir del conocimiento de un experto o puede ser inferido a partir de los datos. Permite establecer relaciones causales y efectuar predicciones.

Redes neuronales

Inspiradas en el modelo biológico, son generalizaciones de modelos estadísticos clásicos. Su novedad radica en el aprendizaje secuencial, el hecho de utilizar transformaciones de las variables originales para la predicción y la no linealidad del modelo. Permite aprender en contextos difíciles, sin precisar la formulación de un modelo concreto. Su principal inconveniente es que para el usuario son una caja negra.

Árboles de decisión

Permiten obtener de forma visual las reglas de decisión bajo las cuales operan los consumidores, a partir de datos históricos almacenados. Su principal ventaja es la facilidad de interpretación.

No supervisados

Descubren patrones y tendencias en los datos sin tener ningún tipo de conocimiento previo acerca de cuales son los patrones buscados.

- Clustering, análisis de enlace, análisis de frecuencia.

Técnicas de clustering

Son técnicas que parten de una medida de proximidad entre individuos y a partir de ahí, buscar los grupos de individuos más parecidos entre sí, según una serie de variables mesuradas.

Conclusiones

La experiencia práctica muestra claramente la aptitud de las técnicas de minería de datos para resolver problemas empresariales. También es clara su aportación para resolver problemas científicos que impliquen el tratamiento de grandes cantidades de datos.

La minería de datos es, en realidad, una prolongación de una práctica estadística de larga tradición, la de análisis de datos. Existe, además, una aportación propia de técnicas específicas de inteligencia artificial, en particular sobre la integración de los algoritmos, la automatización del proceso y la optimización del coste.

A diferencia de la IA, que es una ciencia joven, en estadística se viene aprendiendo de los datos desde hace más de un siglo, la diferencia consiste que ahora existe la potencia de cálculo suficiente para tratar ficheros de datos de forma masiva y automática. Esta es una realidad que cada vez será más habitual. Sin abandonar ninguno de los campos previamente abordados, la estadística ha evolucionado de ocuparse de la contabilidad de los estados a ser la metodología científica de las ciencias experimentales, hasta ser un problema solver para las organizaciones modernas. Es por esta razón el énfasis dado a que los resultados sean accionables.

Por otro lado y en relación a la amplia panoplia de técnicas disponibles, conviene tener claro de que no existe la técnica más inteligente, sino formas inteligentes de utilizar una técnica y que cada uno utiliza de forma inteligente aquello que conoce. También que para la mayoría de problemas no existen diferencias significativas en los resultados obtenidos.

Por todo lo dicho, la minería de datos no es una moda pasajera, sino que se entronca en una vieja tradición estadística y que cada vez más debe servir para hacer más eficiente el funcionamiento de las

organizaciones modernas, ayudar a resolver problemas científicos y ampliar los horizontes de la estadística.

Bibliografía

- Adriaans, P. & Zantige, D. (1996). *Data mining*. Addison-Wesley.
- Aluja,T.&Morineau,A.(1999).*Aprender de los datos, el análisis de componentes principales, una aproximación desde el data mining*. EUB. Barcelona.
- Berry, M. J. A. & Linoff, G. (1997). *Data mining techniques for marketing, sales and customer support*. J. Wiley.
- Lefe ´bure, R. & Venturi, G. (1998). *Le data mining*.
- Eyrolles. McCullagh, P. & Nelder, J. A. (1986). *Generalized Linear Models*.
- Chapman and Hall. Mena, J. (1999). *Data Mining your website*. Digital Press.
- Lebart, L., Salem, A. & Berry, E. (1998). *Exploring Textual Data*, Kluwer, Boston.
- Ripley, B. D. (1996). *Neural Networks and pattern recognition*. Wiley, New York.
- Sarle, W. S. (1994). «Neural Networks and Statistical Models». *Proc*. 9^{th}. *Annual SAS Users Group International Conference*. SAS Institute.
- Sonquist, J. A. & Morgan, J. N. (1964). *The Detection of Interaction Effects*. Ann Ar- bor: Institute for Social Research. University of Michigan.