QÜESTIIÓ, vol. 25, 3, p. 499-508, 2001

APLICACIONES EMPRESARIALES DE DATA MINING

LLUÍS GARRIDO
JOSÉ IGNACIO LATORRE
Universitat de Barcelona*

*El Data Mining, o **extracción de información útil y no evidente de grandes bases de datos**, es una tecnología con un gran potencial para ayudar a las empresas a focalizar sus esfuerzos alrededor de la información importante contenida en sus «data warehouses».*

En este artículo analizaremos las ideas básicas que sustentan el Data Mining y, más concretamente, la utilización de redes neuronales como herramienta estadística avanzada. Presentaremos también dos ejemplos reales de la aplicación de estas técnicas: predicciones bursátiles y predicción de propagación de fuego en cables eléctricos.

Data Mining business applications

Palabras clave: Minería de datos, data mining, redes neuronales

Clasificación AMS (MSC 2000): 62-07, 62-09, 62H30, 68T05

* Departament d'Estructura i Constituents de la Matèria. Universitat de Barcelona. garrido@ecm.ub.es, latorre@ecm.ub.es.

– Recibido en abril de 2001.
– Aceptado en noviembre de 2001.

1. ¿QUÉ ES EL DATA MINING?

La mayoría de compañías tienen una gran cantidad de datos almacenados en sus ordenadores. Estos datos contienen una información que puede ser de gran utilidad para los resultados de la empresa. La gran abundancia de datos o su deficiente estructura puede hacer muy difícil extraer esta información útil. El objetivo del Data Mining [1] es la extracción de forma automática de información relevante, útil y no evidente contenida en dichos datos. Existen tres razones fundamentales por las cuales el Data Mining es una realidad en nuestros días:

- Avances tecnológicos en almacenamiento masivo de datos y CPU.
- Existencia de nuevos algoritmos para extraer información en forma eficiente.
- Existencia de herramientas automáticas que no hacen necesario el ser un experto en estadística, redes neuronales, o algoritmos matemáticos para convertirse en un «DataMiner».

2. ¿QUÉ PUEDE HACER EL DATA MINING?

Una empresa en posesión de unas bases de datos de calidad y tamaño suficiente puede emplear el Data Mining para generar nuevas oportunidades de negocio, dada su capacidad para proporcionar:

- **Predicción automática de comportamientos**.

  Generalmente se trata de problemas de clasificación. como ejemplo podemos citar el marketing dirigido. Data Mining usa los resultados de campañas de marketing realizadas anteriormente para identificar el perfil de los clientes que son más propensos a comprar el producto y de este modo permitirnos substituir el correo masivo por el correo dirigido.

- **Predicción automática de tendencias**.

  Basándonos en base de datos históricas, Data Mining creará un modelo para predecir las tendencias. Como ejemplos podemos citar la predicción de ventas en el futuro o la predicción en mercados de capitales.

- **Descubrimiento automático de comportamientos desconocidos anteriormente**.

  Las herramientas de Data Mining de visualización y clustering, permiten «ver» nuestros datos desde una perspectiva distinta y por ello descubrir nuevas relaciones entre ellos.

3. ¿CÓMO HACER DE DATA MINING? TÉCNICAS

Todo proyecto de Data Mining se desarrolla aplicando ciertas técnicas de especial interés en este campo. Las técnicas más utilizadas son:

- **Redes Neuronales**. Son modelos no-lineales inspirados en las redes de neuronas biológicas y se usan generalmente en problemas de clasificación y predicción. Discutiremos su estructura con un poco más de detalle en los ejemplos.
- **Árboles de decisión**. Son estructuras en forma de árbol que representan conjuntos de decisiones capaces de generar reglas para la clasificación de los datos.
- **Algoritmos genéticos**. Son modelos inspirados en la evolución de las especies y que se aplican generalmente en problemas de optimización. Permiten incluir fácilmente ligaduras complicadas que limitan la solución a un problema.
- **Clustering**. Métodos de agrupación de datos que nos permiten clasificar los datos por su similitud entre ellos. Son utilizadas con frecuencia para entender los grupos naturales de clientes en empresas o bancos.

4. METODOLOGÍA DE DATA MINING

Todo proyecto de Data Mining tiene unas fases bien definidas que van desde la definición del problema hasta la ejecución y evaluación del modelo, pasando por el estudio de los datos y la creación de dicho modelo. Dichas fases quedarán ilustradas en los dos ejemplos que se darán a continuación.

5. REDES NEURONALES

Una red neuronal artificial (ver por ejemplo [2, 3] para una introducción) es un modelo de computación inspirado en nuestros conocimientos sobre neurociencia, es decir, el estudio de las neuronas de nuestro sistema nervioso, aunque sin tratar de ser biológicamente realistas en detalle. En los últimos años estos modelos han experimentado un gran desarrollo gracias al descubrimiento de su excelente comportamiento en problemas de reconocimiento de patrones, predicción y clasificación, entre otros. Las redes neuronales artificiales son mecanismos matemáticos que aprenden a reconocer o clasificar patrones y, tal como lo hace nuestro propio cerebro, dicho aprendizaje no descansa sobre un modelo preconcebido sino que busca las correlaciones existentes entre las variables del problema que se está estudiando.

Para los ejemplos que seguirá hemos escogido redes neuronales multicapa entrenadas mediante el algoritmo de la back-propagation [4, 5]. Una arquitectura típica de di-

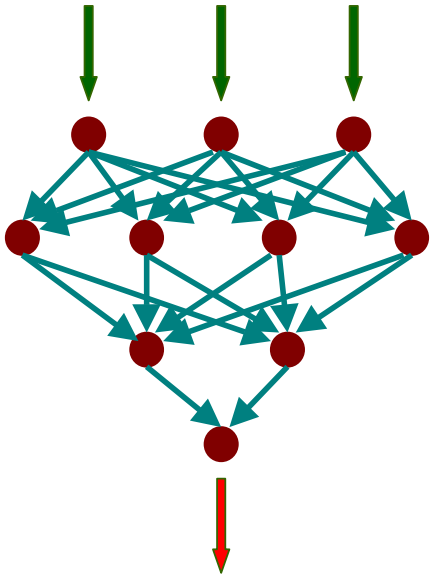

Figura 1. Esquema de una red neuronal multicapa sencilla.

chas redes es similar a la representada en la figura 1. Las neuronas (o unidades) están agrupadas en capas. Existe una primera capa por la que entra la información, una o varias capas ocultas que la procesan, y una capa de salida que proporciona los resultados de la red. Durante la fase de entrenamiento se le presenta a la red un conjunto de posibles valores de entrada juntamente con sus resultados de salida deseados, y el algoritmo de back-propagation busca automáticamente las correlaciones entre dichas entradas y sus salidas respectivas. Cuando el aprendizaje ha finalizado, la red puede ser aplicada sobre datos que no ha visto previamente, realizando sus propias predicciones.

En las redes neuronales multicapa se asocia un peso $W_{ij}^{(l)}$ a la conexión (sinapsis) existente entre la unidad j de la capa $l-1$ y la unidad i de la capa l. La salida de cada unidad $I_i^{(l)}$ se evalua utilizando la expresión

$$I_i^{(l)} = F\left(\sum_j W_{ij}^{(l)} I_j^{(l-1)} + B_i^{(l)}\right),$$

donde $B_i^{(l)}$ es el llamado umbral de la unidad i de la capa l. La función F se conoce como función de activación, y habitualmente se escoge como la función identidad ($F(x) = x$) o como una función sigmoidal ($F(x) = 1/(1+e^{-x})$). En este último caso es conveniente escalar las entradas y las salidas deseadas entre los valores 0 y 1.

Para entrenar una red neuronal con L capas se necesita un conjunto de datos de entrenamiento $\left\{(\mathbf{I}^{(L)\mu}, \mathbf{O}^{(L)\mu}), \mu = 1, \ldots, N\right\}$, donde $\mathbf{O}^{(L)\mu}$ representa la salida deseada correspondiente a la entrada $\mathbf{I}^{(0)\mu}$. El error de la red para estos datos es

$$E = \frac{1}{2}\sum_{\mu=1}^{N}\sum_{i}\left(I_i^{(L)\mu} - O_i^{(L)\mu}\right)^2,$$

donde $\mathbf{I}^{(L)\mu}$ representa la salida de la red después de presentarle una entrada $\mathbf{I}^{(0)\mu}$. La back-propagation intenta minimizar este error modificando la intensidad de los pesos y de los umbrales de forma iterativa, en un proceso conocido como de entrenamiento de la red neuronal. Existen otros métodos de entrenamiento, pero la back-propagation es, con diferencia, el más utilizado. De hecho, la back-propagation es un caso particular del bien conocido algoritmo de descenso por el gradiente, que está basado en una modificación de los pesos según la regla

$$W_{ij}^{(l)} = -\alpha\frac{\partial E}{\partial W_{ij}^{(l)}}$$

Es importante tener presente que la aplicación de las redes neuronales no se reduce únicamente a su entrenamiento con los datos disponibles: uno de los principales problemas es escoger previamente las variables más relevantes, ya que un exceso de variables puede introducir ruido que oculte las más significativas, mientras que su defecto puede provocar una falta de información.

6. PREDICCIONES BURSÁTILES

La aplicación de redes neuronales a la predicción de series temporales ha atraído la atención de mucha gente relacionada con los mercados financieros de todo el mundo. Sin embargo, el hecho de que la evolución de los mercados dependa de multitud de variables, muchas de las cuales son difíciles de cuantificar, hace que sea complicado encontrar situaciones en las que únicamente un análisis numérico de los datos históricos permita realizar buenas predicciones.

Como aplicación práctica de este tipo de redes neuronales, consideraremos el análisis de una acción en la bolsa noruega. Seguiremos paso a paso la metodología básica de la aplicación de esta técnica.

- *Definición del problema*

  Deseamos conocer la evolución de la acción PGS a dos días. El primer paso requiere decidir qué variables serán alimentadas a la red neuronal. Hemos optado por considerar como variables de entrada: la propia serie PGS, el índice Nasdaq, la cotización de

la misma empresa en el mercado americano (PGSUS), el índice de la bolsa noruega y los tipos de interés en Estados Unidos.

- *Preprocesamiento de las variables*

  La red neuronal podrá extraer información útil si las variables que la alimentan están correctamente preprocesadas. Hemos decidido considerar transformaciones de las variables del tipo considerado en el chartismo (medias móviles, ROC, ...) y en otras técnicas de análisis de datos (Gumbel,...)

- *Entrenamiento de las redes*

  Las redes neuronales son entrenadas evitando sobreentrenamiento y falta de generalización. Es conveniente considerar un entrenamiento que corresponda a mostrar cada patrón unos pocos miles de veces a la red neuronal. El entrenamiento breve no logra captar la ley que subyace al mercado. Un entrenamiento demasiado largo hace que la red aprenda los errores (no la ley general) de cada patrón. Una forma de evitar sobreentrenamiento es construir modelos alternativos con redes neuronales pequeñas. Una red pequeña no tiene grados de libertad suficientes para llegar a estar sobreentrenada.

- *Evaluación del modelo*

  Una vez que nuestra red neuronal es entrenada disponemos de un modelo de predicción. Podemos a continuación rastrear el éxito o fracaso de este modelo en los últimos meses. Es importante lograr que nuestro modelo además de ser rentable sea robusto frente pequeños cambios de variables o de estructura de la red neuronal. El modelo debe también describir con igual éxito tanto las subidas como las bajadas del mercado. Por último, el modelo debe proporcionar resultados similares a lo largo del tiempo. Si cualquiera de estos requisitos no se cumple podemos fácilmente haber construido un modelo que tendrá pobres resultados.

- *Definición de estrategia de actuación*

  El modelo neuronal debe considerarse como una herramienta de predicción que se complementa con otros métodos alternativos. Es muy posible que no todos los datos del mercado queden reflejados en las pocas series que alimentan a la red. Una correcta estrategia de actuación deberá contar con los condicionantes de la agresividad del inversor, de su liquidez, etc.

Los resultados obtenidos por las redes neuronales entrenadas siguiendo los pasos anteriores ofrecen resultados muy satisfactorios. Sobre un periodo de seis meses se logran beneficios de un 20%, ignorando comisiones de compra. Los beneficios se obtienen tanto en operaciones de compra como de venta (en el mercado de futuros).

Este ejemplo bursátil se ha implementado con la herramienta Stocknn (http://www.aernsoft.com) desarrollada conjuntamente por las empresas AERN y CAP GEMINI ERNST & YOUNG para la predicción de mercados de capitales con redes neuronales.

7. DATA MINING Y PROPAGACIÓN DE FUEGO

La aplicación de redes neuronales como herramienta de Data Mining tiene un vasto campo de aplicación. Presentamos un segundo ejemplo del empleo de redes neuronales para la predicción de la propagación de fuego en cables eléctricos. Como en el caso anterior, deseamos describir los pasos en que se organiza el análisis de este problema.

- *Definición del problema*

  Una severa normativa regula la construcción de cables eléctrico de forma que la propagación de un eventual fuego sea lo más reducida posible. El problema consiste en predecir la propagación de fuego en diferentes cables de nueva construcción a partir de los datos de sus componentes. Las variables utilizadas para alimentar la red neuronal serán, pues, las características geométricas del cable y las propiedades químicas de sus componentes.

- *Selección de variables*

  En este problema el número de variables es muy elevado. Cada material tiene muchas propiedades químicas. Se hace necesario realizar una selección de las variables más interesantes. Para ello hemos construido grandes redes neuronales que, una vez entrenadas, permiten establecer un criterio cuantitativo para evaluar la relevancia de cada variable. Un buen criterio es la suma de los valores absolutos de los pesos conectados a una cierta neurona i, p_i:

  $$p_i = \sum_j |W_{ij}^{(1)}|$$

  A mayor sea este peso, mayor será la relevancia de esta variable.

- *Entrenamiento de la red neuronal*

  El entrenamiento de una esta red neuronal sigue los pasos del ejemplo anterior. Debe evitarse el sobreentramiento y la falta de entrenamiento.

- *Construcción de un modelo alternativo*

  Para poder evaluar la bondad del modelo neuronal es preciso disponer de un modelo alternativo. En este caso, hemos optado por construir una regresión mulitlineal a las mismas variables que definen el modelo neuronal. La red debe captar las correlaciones no lineales y, por lo tanto, ser superior al modelo lineal.

- *Evaluación de los modelos*

  Para evaluar los modelos se procede a dividir la base de datos en tres partes: datos de entrenamiento, datos de validación y datos de predicción. La red neuronal y el modelo lineal son entrenados empleando los datos de entrenamiento. Luego se aplican ambos modelos en los datos de validación. Así se puede evaluar si ambos modelos funcionan con corrección y en qué medida la red neuronal es mejor que el modelo lineal. Una vez tenemos una correcta evaluación del modelo, podemos aplicarlo sobre los datos de predicción.

- *Clustering*

  Los cables forman grupos naturales con respecto a su comportamiento frente a la propagación del fuego. Hemos empleado la técnica de Análisis de Componentes Principales (PCA) que permite construir una proyección en dos dimensiones del espacio de variables que definen los cables. Cables próximos en cuanto a su construcción son próximos en su proyección.

Las redes neuronales entrenadas para la propagación de fuego logran correlaciones del orden de .8 entre predicción y realidad sobre el conjunto de validación. El porcentaje de acierto en el criterio de paso de normativa de fuego se halla por encima del 80% de los cables.

Este ejemplo de Data Mining en propagación de fuego se ha implementado con la herramienta QnnM (http://quantium.ecm.ub.es) el grupo Quantium de aplicaciones empresariales de redes neuronales.

8. REFERENCIAS

Cabena, P.; Hadjinian, P.; Stadler, R.; Verhees, J. & Zanasi, A. (1997). *Discovering Data Mining from Concept to Implementation*, Book & Cd edition, (September 1997). [1]

Hertz, J. A.; Krogh, A. & Palmer, R. G. (1991). *Introduction to the theory of neural computation*, Addison-Wesley, Redwood City, California. [2]

Müller, B. & Reinhardt, J. (1991). *Neural networks: an introduction*, Springer-Verlag, Berlin. [3]

Rumelhart, D. E.; Hinton, G. E. & Williams, R. J. (1986). «Learning representations by back-propagating errors», *Nature*, 323, 533. [4]

Werbos, P. (1974). *Beyond regression: new tools for prediction and analysis in the behavioral sciences*, Ph. D. thesis, Harvard University. [5]

ENGLISH SUMMARY

DATA MINING BUSINESS APPLICATIONS

LLUÍS GARRIDO
JOSÉ IGNACIO LATORRE
Universitat de Barcelona*

*Data Mining, can be defined as the **art of extracting non-obvious, useful information from large databases**. This emerging field brings a set of powerful techniques which are of relevance for companies to focus their efforts in taking advantage of their data warehouses.*

In this paper we analyse the basic ideas behind Data Mining and we pay special attention to the use of neural networks as an advanced statistical tool. We also present two real world applications of these techniques to stock market predictions and to the study of fire propagation in cables.

Keywords: Data Mining, Neural Networks

AMS Classification (MSC 2000): 62-07, 62-09, 62H30, 68T05

* Departament d'Estructura i Constituents de la Matèria. Universitat de Barcelona. garrido@ecm.ub.es, latorre@ecm.ub.es.

– Received April 2001.
– Accepted November 2001.

A company may keep large amounts of high quality information which can be analysed using Data Mining techniques to uncover new business opportunities, as for instance

- **Targeted Marketing**

  This is a typical classification problem. Data Mining analyses results from previous marketing efforts to identify customer profiles and their predisposition to buy new products. These techniques are a profitable substitute for massive spamming.

- **Characterization and clustering**

  Data can be analysed to visualize and cluster them. Market segmentation and isolation of relevant categories of customers are better characterize using non-linear data mining techniques.

Data Mining projects are developed using a number of advanced statistical techniques. We here mention a few of them:

- **Neural Networks**

  They provide non-linear models inspired in biological neural networks and they are used in problems of classification and prediction. In the paper we discussed practical examples of how to use them.

- **Decision trees**

  These are tree structures representing sets of decisions that can generate rules for data classification.

- **Genetic algorithms**

  These models are inspired in the evolution of species and are applied to optimisation problems. They allow for easily introducing constraints which limit the solution to a problem.

- **Clustering**

  These methods allow for a classification of data given similarities in their patterns. They are used to understand natural groups of customers in large companies and banks.

From the practitioner point of view Data Mining follows well defined phases. The project must be defined, data must be cleaned, several alternative models must be built and a strict validation must check the goodness of the final result. These examples and techniques are discussed in more detail in the paper.