LA MINERÍA DE DATOS: ANÁLISIS DE BASES DE DATOS EN LA EMPRESA

JOSÉ Mª CARIDAD Y OCERIN
ACADÉMICO CORRESPONDIENTE

Los métodos estadísticos son una herramienta imprescindible en la gestión empresarial. La minería de datos emplea de forma sistemática diversas técnicas de análisis de datos en los procesos de toma de decisiones empresariales utilizando la información oculta en grandes bancos de datos que diariamente se generan en la actividad económica, con posibilidad de aumentar el beneficio, pero también con graves riesgos para preservar la intimidad de las personas.

INTRODUCCIÓN

Los métodos de tratamiento de la información en la empresa se iniciaron hace muchos años con la automatización de los procesos repetitivos y administrativos. Los sistemas informáticos centralizados se difundieron en las décadas de los sesenta y setenta en las grandes corporaciones. La aparición de los mini-ordenadores permitió la incorporación en medianas empresas de procesos automatizados, y, finalmente la difusión masiva de los ordenadores personales en los ochenta y de las redes de comunicación generalizaron el uso de los procesos informáticos y obligaron a cambiar las estructuras centralizadas de los centros de proceso de datos.

Las tecnologías de la información están orientadas hoy día, no sólo a los procesos de tratamiento administrativo, sino también hacia la gestión de datos y el soporte en los procesos de toma de decisiones. La difusión de redes de ordenadores, incluyendo los equipos personales, origina una descentralización de la información que dificulta la integración en su uso en la gestión de la empresa. Por otra parte la aparición de nuevas herramientas está facilitando esta integración y uso más eficiente a través de dos tipos de desarrollos tecnológicos: los denominados *Data Warehouse* (DW) *o almacén de datos, y Data Mining* (DM) *o minería de datos*.

Varios factores han permitido estos desarrollos: la reducción continua de los costes de almacenamiento y proceso de la información, el incremento de la potencia de cálculo a través de varias tecnologías (SMP o *Symmetric Multi Processing*, en el que en un solo sistema varios procesadores se reparten en el trabajo, SMC, cluster o conjunto de ordenadores que comparten los mismos sistemas de almacenamiento de datos, o los MPP o multiprocesadores masivamente paralelos interconectados por canales muy rápidos que permiten considerarlos como un único sistema), y las necesidades derivadas del incremento de productividad y de tratamiento individualizado de la clientela. Además,

en los últimos veinte años han surgido varias herramientas en el campo de software para el acceso a la información; de nuevo es frecuente el uso de siglas como DSS (*Decision Support Systems*, o sistema de soporte a las decisiones), OLAP (*On Line Analytical Processing*, en el que se procesa e integra la información para la gestión a medida que se produce), MDA (Análisis multidimensional de datos), y, las ya citadas DW y DM.

Los sistemas de consulta de una base de datos permiten acceder a registros o subconjuntos de información que cumplen unas determinadas condiciones. Los métodos OLAP responden a preguntas sobre porqué son ciertas algunas afirmaciones, incorporando algunos contrastes estadísticos usuales. Los métodos DM siguen un proceso inductivo orientado a buscar asociaciones no conocidas entre variables o casos, incorporando nuevas hipótesis al trabajo. Por ejemplo, con las primeras herramientas pueden asociarse, en una base de datos bancaria, que el nivel de renta y patrimonio son factores de riesgo, al analizar un crédito; un paquete de DM debe además proponer que existan otras variables, como la edad del peticionario, para decidir. Los sistemas OLAP son útiles en fases iniciales de análisis de datos. Los métodos de DM utilizan modelos estadísticos, métodos de visualización, y técnicas de inteligencia artificial, en los procesos empleados, y, son posibles gracias a la existencia de procesos computacionales intensivos y de lamacenamiento de grandes volúmenes de información. En los sistemas de consulta a bases de datos, el usuario obtiene información que ya existe, mientras que con DM se persigue generar nuevos elementos de información.

LOS CENTROS DE INFORMACIÓN Y LA MINERÍA DE DATOS

En IBM se desarrollaron hace veinticinco años los generadores de informes, y posteriormente los sistemas de consulta (QUERY). Posteriormente los lenguajes SQL se emplearon para realizar consultas a las bases de datos, si bien requieren un personal medianamente especializado. La aparición de los ordenadores personales originó una descentralización de la información lo que introdujo una dificultad adicional en la integración de datos para la toma de decisiones. Los Centros de Información (CI) se crearon para agrupar los datos de uso general y proporcionar ficheros a los ordenadores personales de los usuarios finales, para ser tratados en ellos con las herramientas de ofimática usuales. La falta de visión global de la información en los CI, su dispersión y la existencia de datos redundantes ha originado la aparición de las DW.

El almacén de datos o DW engloba la información de cada área de la empresa destinada a las necesidades de los sistemas de soporte a las decisiones y para la gestión y control del funcionamiento empresarial. Debe incorporar bases de datos integradas y con una terminología normalizada para las distintas aplicaciones (por ejemplo, los campos relativos a cada cliente se identifican siempre de la misma forma). Además las bases de datos deben ser temáticas (orientadas hacia aplicaciones definidas, como la gestión de clientes, proveedores, productos, etc., en lugar de hacia procesos administrativos) e incluir información histórica, es decir, los datos serán series temporales asociadas a cada instante del tiempo, manteniéndose a lo largo de varios años.

Así pues, la información que llega al DW procede de diversas fuentes: los procesos operativos de gestión, integrando o agregando los datos procedan; también se suelen incorporar datos externos (de clientes, de tipo socioeconómicos, etc) y de los ordenadores departamentales y personales.

La centralización y homogeneización de la información está destinada a permitir visiones globales aplicables en las decisiones de gestión y a controlar la fiabilidad de

los datos. Este enfoque origina además una reducción de costes en la obtención de la información y garantiza una mayor calidad de ésta. Por lo tanto, el DW es el proceso de organización de grandes volúmenes de datos generalmente multidimensionales, para facilitar el acceso a la información con fines analíticos.

Las herramientas de acceso al DW son variadas. Los sitemas de consulta SQL y de informes dan paso a otros programas específicos para extraer información. El análisis multidimensional permite al usuario construir informes sin necesidad de conocer la estructura interna de las bases de datos, con unas utilidades para definir los subconjuntos de datos. Por último, el DM o minería de datos, trata de realizar análisis estadísticos y síntesis automática de la información buscando regularidades, dependencias, patrones de comportamiento, grupos de casos, etc., que no son evidentes; es decir, descubrir información útil contenida en los datos.

En los DW es fundamental la estructuración de la meta-información, esto es, la información sobre la información. Entre ésta se incluye los nombre de los campos, el tipo de datos que contienen, sus relaciones con otros, las propiedades de la información, etc. Esta meta información supone un soporte al usuario del DW, facilitándole el acceso y la elaboración de consultas, informes y análisis, pero también debe servir de soporte a los responsables técnicos del DW.

Los gestores de base de datos son elementos fundamentales en los DW. Las bases de datos relacionales estructuran la información en forma de tablas o matrices de casos, para los que se disponen de datos de varias variables o campos. Los sistemas relacionales incluyen varias tablas interrelacionadas, y con los adelantos en los sistemas de almacenamiento y velocidad de los equipos multiprocesadores, son adecuados incluso en consultas que requieren manejar grandes volúmenes de información.

Generalmente los sistemas DW tienen un número limitado de usuarios, pero en sus consultas acceden a un volumen muy elevado de datos; la escalabilidad, o capacidad de crecimiento, tanto de los equipos físicos como de los programas, es un factor importante en el diseño de estos sistemas. El modelo y la arquitectura de datos de un DW deben ajustarse a las funciones y necesidades de la gestión empresarial; así pues, es preciso que se consiga accesibilidad a la información, la cual debe ser uniforme y clara, actualizada y fiable. Su implantación suele ser gradual, y a través de un proyecto a nivel de toda la empresa.

El concepto de minería de datos (DM) representa unas ideas que han venido madurándose a lo largo de muchos años: como recorrer grandes bases de datos para recuperar información conceptual de interés y para inferir nuevas informaciones útiles. No se trata de una simple búsqueda a través de palabras clave o descriptores, pues es frecuente que no se conozca a priori exactamente lo que se busca, o lo que se puede encontrar. Por ejemplo, una cadena de hipermercados puede estar interesadas en tendencias generales o agrupación geográfica de las ventas que no son evidentes en sus operaciones diarias. Las técnicas DM utilizan algoritmos matemáticos y estadísticos, para realizar búsquedas de patrones o comportamientos sistemáticos que pongan de manifiesto interrelaciones entre los datos o que sirvan para predecir comportamientos futuros. Es decir, las técnicas de DM más usuales se orientan a la predicción automática de tendencias y comportamientos, y al descubrimiento de patrones desconocidos existentes en bases de datos, generalmente integrados en un DW.

La minería de datos consiste en diversos conjuntos de procesos analíticos para explorar grandes conjuntos de datos. Los objetivos son diversos: el descubrir pautas de comportamiento o interrelaciones sistemáticas entre variables, generalmente de bases de datos empresariales, construir modelos predictivos, y en general extraer información

no evidente utilizando métodos computacionales intensivos. D. J. Hand (2000), define la minería de datos como el descubrimiento de estructuras interesantes, inesperadas o valiosas, en grandes conjuntos de datos. En los procesos de DM las técnicas de análisis de datos se usan *a posteriori*, es decir una vez que se han recogido los datos, pues estos se obtienen con otros fines, como el emitir una factura en una transacción comercial, o realizar un apunte contable. Generalmente esta recogida de datos se realiza por procedimientos automáticos, y para cumplir unos fines básicos empresariales. La minería de datos es un proceso posterior, destinado a un mejor conocimiento de la información disponible, a aumentar beneficios o ventas, a disminuir pérdidas, es decir con un objetivo distinto al que ha motivado la recogida y almacenamiento de información. No son pues habituales el analizar problemas de diseños muestrales para realizar esta tarea.

Algunos ejemplos de aplicación de las técnicas de DM son las siguientes: localización de un conjunto de consumidores que tienden a rersponder a una campaña de publicidad por correos; predecir los fallidos en créditos al consumo; reducción de errores en los procesos de fabricación; estimación de la audiencia de programas de televisión; determinación de las características de los clientes que originan mayores beneficios. Más adelante se proporcionan algunos ejemplos adicionales.

Las técnicas empleadas en el DM son muy variadas, pues no todas son aplicables en cualquier conjunto de datos. Generalmente la generación de informes y los métodos OLAP ya citados, consistentes en el procesamiento y análisis de datos a medida que éstos se van produciendo, están integrados en los sistemas de minería de datos. El empleo de métodos estadísticos de uso frecuente son los siguientes: decripción uni y multivariante de datos, incluyendo las correspondientes técnicas gráficas, diversos contrastes de hipótesis, modelos de regresión y de regresión logística, análisis discriminante, análisis cluster o de conglomerados, técnicas de reducción de dimensión como el análisis en componentes o en coordenadas principales, o el análisis factorial, series temporales, árboles de decisión, redes neuronales, algoritmos genéticos y otras técnicas estadísticas de visualización y presentación de datos, y de sus interacciones.

En todo caso, la aplicación de técnias de DM en la empresa, o en una institución, requiere un conocimiento profundo del negocio así como de los datos que existen en las bases de datos corporativos, y también es preciso entender los métodos analíticos empleados y sus limitaciones. Las posibilidades de obtener información de interés se acrecientan si se conoce a fondo el objetivo de recogida de datos, el tipo de variables que se miden, su calidad y redundancia, y la familiaridad con el los objetivos generales de la empresa en relación a la información acumulada. En las bases de datos corporativas está la información básica que se procesa en cualquier estudio de DM; sin embago, es conveniente extraer los datos que interesan en una base de datos específica, para no interferir en los procesos administrativos ordinarios. Así pues, si es necesario alterar algún dato, como por ejemplo, una corrección de un dato anormal, o la generación de nuevas variables, o la imputación de datos que faltan, será más práctico disponer de una base de datos en la que poder operar y realizar simulaciones, sin peligro de alterar datos históricos. Además de las bases de datos corporativas no suelen tener estructura adecuada para ser incorporadas a un proceso de DM. Una vez generada la base de datos específica, conviene realizar un control de calidad sobre éstos, con la correspondiente depuración, así como añadirles los elementos de meta-información necesarios para poder aplicar métodos estadísticos o de DM. Es frecuente que sea necesario añadir datos adicionales, que provienen de otras fuentes, generar variables a partir de las existentes y realizar diversos procesos de agregación con todos o con parte de los datos. También es necesario especificar las restricciones y protocolos de uso de los datos para mantener la

confidencialidad y privacidad pertinente. Una vez preparadas las bases de datos se podrá iniciar el proceso de análisis, tanto descriptivo como la elaboración de modelos y la obtención de relacciones, segmentaciones, y otras técnicas estadísticas.

Al realizar un estudio, generalmente no se trata de poder predecir los datos de una o de varias variables contenidas en la base de datos, sino obtener resultados que puedan aplicarse a nuevos datos no disponibles todavía. Para ello hay que realizar un muestreo en la base de datos y probar los modelos estimados con datos no incluidos en las muestras seleccionadas. Posteriormente, y a medida que se generan más datos, hay que realizar un seguimiento de los modelos obtenidos, para actualizar sus coeficientes, y para comprobar que siguen manteniendo poder predictivo.

No hay que olvidar que el uso de la DM tiene como objetivo la toma de decisiones de gestión, y, por lo tanto sus resultados deben poder aplicarse por el gerente o decisor correspondiente. Algunos programas incorporan herramientas para transformar estos resultados en medidas económicas, ratios, índices, y modelos de gestión facilitando así la aplicabilidad, e incluso, realizando una valoración de los resultados en función de los beneficios que origina la aplicación del modelo, o de la disminución de costes.

El éxito en la aplicación de la DM depende de dos factores: el planteamiento claro del problema y de los objetivos, y la disponibilidad de datos adecuados. La calidad y fiabilidad de la información es importante, pues numerosas técnicas estadísticas son muy sensibles a la presencia de datos anormales o no representativos.

En principio la minería de datos se aplica sobre bases de datos que se han obtenido sin ningún diseño muestral. por ejemplo las ventas de una empresa un día constituyen un colectivo, aunque, a veces, puede considerarse como una muestra de una población mayor, si se van a realizar inferencias sobre las ventas en días sucesivos, Algunas técnicas estadísticas requieren realizar un muestreo en una base de datos, especialmente si se trata de elaborar modelos, que posteriormente hay que validar con casos no empleados para estimarlos. Pero generalemente las decisiones se aplican a los casos, por lo que hay que emplear toda la información disponible realizando un análisis descriptivo de la población disponible.

Las técnicas estadísticas multidimensionales incorporadas a los programas de DM suelen agruparse en varios bloques. En primer lugar los **métodos de clasificación**, como el análisis de conglomerados (*cluster*), cuyo objetivo es descubrir conjuntos de casos (clientes, ventas, productos...) o de variables que son similares y que se agrupan o tienen características similares. Esta segmentación define tipologías o clases de elementos "parecidos". Así, si dos elementos (casos o variables) se definen mediante la observación de p características, es necesario definir una medida de distancia o similaridad entre ellos. Por ejemplo, la distancia euclídea

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{p}(x_i - y_i)^2}$$

es un ejemplo, aunque no es el más utilizado. Existen otras medidas de distancia para considerar problemas derivados de las distintas escalas de medida cada una de las diferentes características. Para datos no numéricos también se definen diversas medidas de similitud o distancia. Finalmente los m elementos se clasifican en función de la correspondiente matriz de distancias **D**; los resultados se suelen representar gráficamente en forma de árbol o dendograma, para poder visualizar las proximidades. Otras técnicas para poder realizar estas clasificaciones se basan en métodos estadísticos, como las redes neuronales o algoritmos genéticos, que se han desarrollado en el ámbito de la

inteligencia artificial. El análisis discriminante y los modelos de variable respuesta cualitativa son otras técnicas de clasificación; la diferencia fundamental con el análisis cluster estriba en que las clases o agrupaciones son conocidas a priori. Es decir, no se intenta descubrir conglomerados, sino usar casos previamente clasificados para definir una reglas de clasificación. El uso de funciones discriminantes o cuadráticas puede ser empleado, o alternativamente, el empleo de redes neuronales permite determinar reglas de clasificación de tipo no paramétrica, que proporcionan mejores resultados si la separación entre clases no es lineal. El reconocimiento de patrones se usa para asociar estructuras de datos a unas configuraciones predeterminadas, para catalogar casos. En otros ámbitos se han empleado para el reconocimiento óptico de caracteres, análisis gramatical de textos y sistemas de visión artificial. También es posible establecer una clasificación mediante la definición de un conjunto de reglas independientes. No precisa establecer una división jerarquica con un esquema en árbol, aunque las reglas pueden dar lugar a situaciones contradictorias. En el proceso de clasificación es usual atribuir un nivel de confianza a cada regla.

Otros métodos estadísticos de uso frecuente son los orientados a la **reducción de la dimensión**; generalmente los datos contenidos en una base de datos son de naturaleza multidimensional. Cada elemento (registro, caso,...) lleva asociado varias variables. Es muy frecuente que la información sea parcialmente redundante, y que pueda represetarse en un subespacio de dimensión menor con una pérdida mínima de información. De esta forma es posible, a veces, interpretar el fenómeno objeto de estudio en un contexto más simple y obtener así información útil, difícil de conseguir sobre los datos originales. El análisis en componentes y el de coordenadas principales persiguen estos fines. Dado un conjunto de datos en los que se miden p variables numéricas, $x_1, x_2, ..., x_p$, es frecuente que la información que contienen sea en parte redundante, debido a la existencia de interrelaciones entre los datos. Si las relaciones (no exactas) son lineales, las variables están correlacionadas. El análisis en componentes se basa en realizar la transformación lineal

$$\vec{\mathbf{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} a_{11}\ a_{12} \cdots a_{1p} \\ a_{21}\ a_{22} \cdots a_{2p} \\ \cdots\cdots\cdots\cdots \\ a_{p1}\ a_{p2} \cdots a_{pp} \end{pmatrix} \begin{pmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \\ \cdots\cdots \\ x_p - \bar{x}_p \end{pmatrix} = \mathbf{A}(\vec{\mathbf{x}} - \vec{\bar{\mathbf{x}}})$$

de tal forma que las variables $y_1, y_2, ..., y_p$, denominadas componentes principales sea incorreladas y la varianza de y_1 sea máxima, la de y_2 sea máxima condicionada a estar incorrelada con y_1, así sucesivamente. En el caso que la varianza agregada de las primeras componentes sea un porcentaje elevado de la varianza total agregada de las variables $x_1, x_2, ..., x_p$, es posibles usar estas primeras componentes en sustitución de las p variables originales, con una pérdida mínima de información. Los modelos de análisis factorial persiguen representar un conjunto de variables observadas $x_1, x_2, ..., x_p$, en función de unas variables no observables denominadas factores. Es usual suponer que existe un número $m<p$ de factores comunes, $f_1, f_2, ..., f_m$, que influyen sobre todas las variables, y otros factores específicos $e_1, e_2, ..., e_p$, de cada variable, dando lugar a un modelo de la forma.

$$\vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} = \begin{pmatrix} a_{11} \cdots a_{1m} \\ a_{21} \cdots a_{2m} \\ \cdots\cdots\cdots \\ a_{p1} \cdots a_{pm} \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_p \end{pmatrix} = \mathbf{A}\vec{\mathbf{f}} + \vec{\boldsymbol{\varepsilon}}$$

Al estimar el modelo se usan los datos de las p variables x, y se trata de analizar si efectivamente el modelo representa adecuadamente a las observaciones. Existen varias técnicas adicionales de reducción de dimensión como el análisis canónico, el análisis en coordenadas principales o los escalogramas múltiples. En todos los casos se trata de intentar obtener una representación simplificada de la realidad, que no es evidente a partir de los datos originales.

Otras técnicas persiguen buscar **asociación** o relaciones entre sucesos o casos aparentemente independientes, y como la ocurrencia de un suceso puede servir para predecir la de otros. Este tipo de herramientas se utilizan, por ejemplo, en el análisis de los hábitos de compra de los clientes de una empresa. Los patrones secuenciales están orientados hacia la detección de las asociaciones temporales entre los distintos sucesos. Los perfiles se generan cuando se produce una situación anómala o relacionada con alguno de los objetivos empresariales, y en minería de datos, es más frecuente asociar medidas o puntuaciones de similaridad o lejanía a los datos, que realizar contrastes estadísticos para detectar perfiles o diferencias significativas. No hay que olvidar que las técnicas asociadas a la obtención de perfiles se aplican a casos concretos, y no a conjuntos o muestras.

Para **predecir** la evolución de una o varias magnitudes se utilizan diversas técnicas. Algunas se basan en prospecciones y encuestas a expertos, pero la situación más usual es el empleo de modelos dinámicos. Una serie temporal es una secuencia de observaciones $y_1, y_2, ..., y_t, ..., y_n$, en distintos instantes del tiempo. La serie puede ser unidimensional o multivariante, y pueden existir o no variables causales de su evolución. Los casos más usuales de predicción es cuando se dispone de una serie univariante, y_t, y se trata de construir un modelo para realizar predicciones. Existen numerosas técnicas de análisis de series, como por ejemplo, la basada en modelos ARIMA, los modelos clásicos de alisado exponencial, medidas móviles y los modelos de regresión. En todas ellas se persigue realizar una secuencia de transformaciones sobre la serie hasta dejarla reducida a una serie residual en la que no se observan relaciones de los datos actuales con su pasado. En la metodología de Box y Jenkins es habitual tras haber depurado y homogeneizado los datos, eliminar tendencias en variabilidad, en su medida (tendencia y ciclos) hasta llegar a una serie estacionaria para las que se usan modelos ARMA. Los métodos clásicos tratan sobre la desagregación de las serie en diversas componentes (tendencias, ciclos, componentes irregulares) y construir modelos temporales para éstas. Las series multivariantes pueden representarse con distintos tipos de modelos: VAR, espacio de estados, MARMA y modelos econométricos multiecuacionales. Los dos primeros son más fáciles de aplicar y junto con los métodos univariantes son los más usados para realizar predicciones a corto plazo.

La **simulación** estocástica es una técnica utilizada en aplicaciones económicas e industriales. Consisten en utilizar un modelo teórico que se comporte como un sistema real, y generar datos aleatorios compatibles con este sistema. De esta forma es posible obtener un gran conjunto de datos sin necesidad de realizar un costoso trabajo de campo.

Los métodos de **optimización** se encuadran dentro de la Investigación Operativa. Incluyen técnicas muy diversas y de gran utilidad en la empresa: programación lineal, no lineal y dinámica, problemas de transporte, asignación y gestión de inventarios, teoría de colas y otras.

Los árboles de decisión son herramientas empleadas para la clasificación, generalmente basadas en reglas formuladas sobre variables incluidas en las bases de datos. La finalidad de los árboles de decisión es llegar a una clasificación de los casos, con el objetivo de adoptar una decisión. La estructura de un árbol de decisión es similar

al árbol de ficheros de una unidad de disco: la raíz representa la base de datos analizada, y se van generando unas ramas usando diversos criterios. Si al descender un nivel, se obtienen dos ramas, el árbol se dice binario, aunque no son infrecuentes los árboles en los que de cada nodo parten más de dos ramas. Las ramas terminales llevan aparejadas decisiones, o modelos predictivos de las variables objeto de estudio. El programa CART genera árboles binarios y un conjunto de reglas aplicables para clasificar nuevos datos. CHAID es el paquete incluido en SPSS; al construir los grupos trata de maximizar una medida de distancia entre los grupos que se forman. Si la variable de decisión es categórica, el árbol de decisión genera una clasificación, mientras que si es numérica, se estima un modelo de regresión para los casos correspondientes a las ramas terminales. En el caso que el árbol sea excesivamente grande, surgen problemas en la interpretación de los resultados, y, a veces equivale a una sobreparametrización: es decir, se pierde capacidad predictiva para clasificar nuevos casos. Para evitar esta situación, se incluyen reglas de parada basadas en el número máximo de niveles que puede tener el árbol, o sobre el número mínimo de casos en las ramas terminales. A posteriori se usan técnicas de eliminación de ramas, para mejorar la interpretabilidad. J. H. Friedman, uno de los creadores de CART propone otra metodología alternativa, para eliminar algunos de los incovenientes de los árboles de decisión, como el hecho de coordinar las ramas a las divisiones previas, o la dificultad en la interpretación de las interacciones entre variables. El programa MARS incluye predictores discontinuos en la elaboración de las ramas, y elimina la dependencia de cada clasificación respecto a las anteriores, aunque se pierde la estructura gráfica; busca determinar variables causales y sus interacciones mediante modelos no lineales.

Los paquetes comerciales de DM tratan de acercarse a los usuarios finales, integrándose en las necesidades de la empresa. Una de las herramientas que incorporan son facilidades para exportar modelos estimados a conjuntos de datos sobre los que se deben aplicar; algunas tecnologías, como los procesos OLE deben facilitar esta línea en el futuro. También deben tender a integrar procesos de modelización para problemas específicos que se presentan en la empresa, con objeto de facilitar la utilización para analistas de escasa formación estadística. Así pues es necesario poder realizar directamente algunos cálculos financieros con los resultados de los análisis de datos y con los modelos construidos.

OTRAS TÉCNICAS COMPUTACIONALES EN MINERÍA DE DATOS

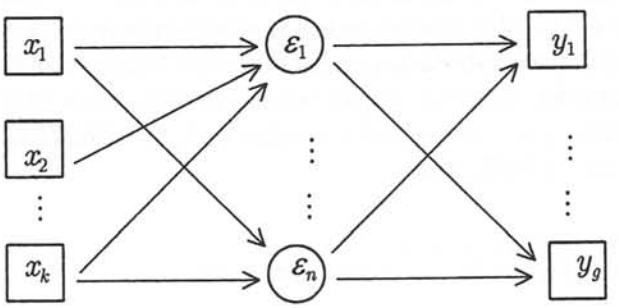
Las técnicas estadísticas y de investigación operativa son las herramientas básicas que se han empleado en DM. Es preciso que el usuario conozca sus fundamentos (sin necesidad de una profundización matemática importante), y, no menos importante es ser consciente de sus limitaciones y ámbito de aplicación. En los últimos años se han incorporado, tanto a los paquetes estadísticos tradicionales, como a los sistemas de DM unas técnicas estadísticas y algoritmos que resultan especialmente útiles al analizar grandes volúmenes de información.

Los **sistemas expertos** consisten en unos programas de ordenador que permiten simular el comportamiento de expertos humanos en los procesos de toma de decisiones. Para ello es necesario formar una base de datos de conocimiento, tras una formalización de ésta. Sobre esta información se definen una serie de reglas de comportamiento que regulan el funcionamiento del programa. El motor de inferencia gestiona las distintas preguntas y entrada de información, y, a medida que es utilizado un sistema experto, se va completando la base de datos de conocimientos con una nueva información, con el

objeto que el sistema se ajuste a un proceso de aprendizaje. Existen diversos tipos de sistemas expertos, y algunos utilizan reglas no deterministas utilizando motores probabilísticos.

La **lógica difusa** es empleada por numerosos modelos y productos industriales para representar situaciones en las que los problemas de clasificación están afectados de incertidumbre. Así en un conjunto borroso, la pertenencia de un elemento a él se formula en términos probabilísticos.

Las redes neuronales constituyen una forma alternativa de representación de modelos y técnicas estadísticas, como la regresión uin y multivariante, la clasificación y otros métodos. Aunque sus raíces están en un intento de representar el funcionamiento del cerebro, los modelos resultantes pueden formalizarse de manera similar a otras técnicas estadísticas. En muchos sistemas se dispone de una información causal, es decir, unas variables $x_1, x_2, ..., x_k$, que influyen sobre un conjunto de variables de salida, $y_1, y_2, ..., y_g$ (dependientes o endógenas), pero de tal forma que las interdependencias no son lineales ni evidentes. Para elaborar un modelo se definen uno o más conjuntos (capas) de variables no observables (neuronas) que sirven de elemento transmisor de la información entre las variables x y las y.



En la figura anterior se representa una red con n neuronas en una sola capa (oculta). Cuando una neurona recibe una información de entrada (el valor de una variable) se activa y produce una salida función de unos pesos (similares a los coeficientes de regresión). Al llegar estas señales a las variables dependientes se obtienen uno valores estimados $\hat{y}_1, \hat{y}_2, ..., \hat{y}_g$; los pesos se van modificando (estimando) hasta conseguir unos errores mínimos: este proceso interactivo se denomina adiestramiento de la red. De esta forma se llega a estimar un algoritmo que permite predecir las variables y en función de las x, sin necesidad de especificar funciones paramétricas como en los modelos de regresión. Si se trata de resolver un porblema de clasificación, como en el caso del análisis discriminante, basta usar una variable de salida no numérica asociada a cada una de las categorías o subpoblaciones. Las redes neuronales proporcionan buenos resultados en problemas no lineales cuando se dispone de un número elevado de datos.

No obstante, la mayor parte de la información obtenida de una base de datos empresarial se obtiene con técnicas descriptivas elementales, en las que se incorporan algunos índices o cálculos fianancieros para facilitar la compresión de los resultados, y su aplicabilidad. Por ejemplo, supóngase que se desea investigar la posibilidad de incrementar las ventas de tres productos, analizando como se comportan los clientes de una empresa respecto a la compra simultánea de dos o más productos distintos. Para ello se extrae de la base de datos de ventas la variable

$$X = (A, B, C, N)$$

en la que las componentes A, B y C indican si se ha producido una compra del primer

producto, del segundo producto o del tercero, respectivamente, y la variable N representa la compra de otros productos de la empresa, salvo los tres considerados. Así pues, al cruzar esta variable consigo misma se obtiene la siguiente tabla de frecuencias de compras, para una muestra de 1000 clientes.

	A	B	C	Total
A	30	15	10	50
B		60	10	80
C			5	20
N				880

Es decir, que en 30 casos solo se ha comprado A, en 15 se han adquirido simultáneamente A y B, en 10 A y C. Como el número total de ventas de A es igual a 50, quiere decir que en 5 ocasiones se han comprado simultáneamente tres productos, de los anteriores. El producto B se ha comprado aisladamente en 60 ocasiones, y en 10 conjuntamente con el C, y como con A se compró en 15 ocasiones, el total de ventas de B son 80. Finalmente, el producto de C se ha comprado en 20 ocasiones, de las que 5 han sido adquirir otros productos, y otros productos (N) se han adquirido sin comprar A, B o C, en 880 ocasiones. A partir de estos datos es posible construir unos índices o proporciones para evaluar las posibilidades de compra conjunta de varios de estos tres productos. Los índices son estimaciones de las probabilidades condicionadas de varios sucesos: si un cliente ha comprado el producto A, la estimación de la probabilidad que compre B es del 20%, y si se sabe que ha comprado A, la estimación de la probabilidad que compre B es del 19.75%, pues

$$\text{freq}(B|A) = \text{freq}(A \cap B)/\text{freq}(A) = 0.015/0.050 = 0.30$$
$$\text{freq}(A|B) = \text{freq}(A \cap B)/\text{freq}(B) = 0.015/0.080 = 0.1875$$

Análogamente se obtienen

$$\text{freq}(C|A) = \text{freq}(A \cap C)/\text{freq}(A) = 0.010/0.050 = 0.2$$
$$\text{freq}(A|C) = \text{freq}(A \cap C)/\text{freq}(C) = 0.010/0.020 = 0.5$$
$$\text{freq}(B|C) = \text{freq}(B \cap C)/\text{freq}(C) = 0.010/0.020 = 0.5$$
$$\text{freq}(C|B) = \text{freq}(B \cap C)/\text{freq}(B) = 0.010/0.080 = 0.1275$$

No obstante, como los tres productos no son demandados en la misma poporción, cabe calcular los ratios entre las frecuencias condicionales anteriores, y las frecuencias de compras de cada uno de los productos.

$$\text{freq}(B|A)/\text{freq}(B) = 0.3/0.08 = 3.75$$
$$\text{freq}(C|A)/\text{freq}(C) = 0.2/0.02 = 10$$
$$\text{freq}(A|B)/\text{freq}(A) = 0.1875/0.05 = 3.75$$
$$\text{freq}(C|B)/\text{freq}(C) = 0.1275/0.02 = 6.375$$
$$\text{freq}(A|C)/\text{freq}(A) = 0.01/0.05 = 0.2$$
$$\text{freq}(B|C)/\text{freq}(B) = 0.01/0.08 = 0.125$$

Así pues, si un cliente ha comprado el producto A, es más probable que compre también el B en lugar del C, siendo la razón de estas probabilidades estimadas 0.3/0.2= 1.5. Como el producto B se compra cuatro veces más que el C, será más eficiente promover la compra del producto B para los clientes del A. Además como el producto B es comprado más frecuentemente que el A, para los clientes de B es más probable que también lo sean del A en relación al C, siendo la razón de las probabilidades estimadas 0.1875/0.1275= 1.47. De igual forma se realizan otros análisis, y pueden incorporarse

árboles de decisión o programas de simulación adicionales, utilizando beneficios esperados por la venta de cada unidad de uno de los productos. Si se han comprado dos de los tres productos, también es inmediato realizar el análisis de las posibles ventas cruzadas.Es habitual realizar consultas para identificar en una base de datos todos los productos en los que la frecuencia de compras condicionales sean superior a un determinado valor, incorporando posiblemente otras condiciones, como por ejemplo, algunas relativas a la difusión del producto entre los clientes de la empresa.

En los paquetes DM es habitual utilizar una terminología que tiende a acercar al usuario final, generalmente un analista de una empresa. Sin embargo, y como es bastante frecuente en algunas aplicaciones estadísticas en ciencias sociales y experimentales, se reinventan conceptos conocidos y bien definidos en textos elementales de Estadística. Por ejemplo, a la frecuencias condicionadas anteriores a veces se les denomina "confianza", o a la frecuencia de compra de un producto "soporte". Es claro que todo ello solo introduce un elemento de confusión conceptual en el uso de herramientas cuya utilidad es evidente. La causa sin duda se debe a un intento de dar apariencia de complejo y original a técnicas bien conocidas. Aunque se pretende que las herramientas de DM sean intuitivas y fáciles de utilizar para un usuario no estadístico; es evidente que el uso de conceptos y técnicas cuantitativas solo se puede realizar con una compresión mínima de éstas, sin necesidad de profundización matemática, pero, ineludiblemente hay que conocer el campo de aplicación de cada técnica, sus limitaciones, y como interpretar las medidas obtenidas.

Además, hay que considerar que para realizar inferencias sobre una base de datos, es necesario emplear un diseño muestral adecuado, que permita una extrapolación fiable al aplicar los resultados del análisis de nuevos casos. Es pues necesario disponer de unos conocimientos básicos de las técnicas de muestreo en poblaciones finitas, y los conceptos asociados a ellas.

APLICACIONES DE LA MINERÍA DE DATOS

Son numerosas las aplicaciones de la minería de datos en el ámbito de la empresa. En primer lugar cabe citar las cuestiones relacionadas con la gestión de todas las fases del ciclo de clientes: desde la adquisición de nuevos clientes y el mantenimiento o retención de la clientela, e identificación de los perfiles de los buenos clientes, hasta el diseño de estrategias para aumentar los ingresos de los clientes habituales. Por ejemplo, se pueden investigar que tipos de clientes no han comprado determinado producto que sin embargo es demandado por otros clientes aparentemente similares. También se pueden investigar que circunstancias se han dado en clientes perdidos, para poder formular predicciones sobre los clientes que se pueden perder en el futuro.

Así mismo se emplean técnicas de DM en análisis de campañas comerciales, proveedores, e incluso en la gestión de inventarios. En los procesos de realización de encuestas, son bien conocidos los estudios de aquellas que tienden a contestar con más frecuencia a las solicitudes de información para conseguir una tasa de respuesta más elevada.

Otro de los sectores en los que se han empleado con asiduidad el DM es en el ámbito financiero. Las compañías de seguros y los bancos lo emplean para analizar el perfil de los tomadores de seguros y operaciones crediticias, y en el estudio de los perfiles asociados a las operaciones más rentables. Las tarjetas de crédito, y su uso fraudulento son objeto de análisis detallado.

Actualmete algunas administraciones públicas están usando las técnicas de minería

de datos para analizar las pautas que llevan a la proliferación de incendios forestales y a catástrofes medioambientales.

Los métodos de DM se diferencian de otras tecnologías de análisis de la información, como las ya citadas, OLAP, las cuales permiten realizar análisis descriptivos, incluso multidimensionales, utilizando herramientas de consulta eficientes, para acceder a datos interdepartamentales. Los resultados finales pueden ser muy simples (tablas de frecuencia y de contingencia, descripción analítica de datos, e incluso métodos más complejos para realizar predicciones o detectar datos anormales). La minería de datos van más allá, pues pueden operar sobre resultados obtenidos con OLAP, el cual actúa de forma deductiva, mientras que el DM incluye procedimientos inductivos. No se limita a verificar hipótesis o tendencias, sino que el DM pretende descubrir nuevas pautas y tendencias de los datos. Por ejemplo, con una herramienta de tipo OLAP puede asociar el riesgo de mal uso de una tarjeta de crédito al nivel de renta, mientras que un paquete de DM puede además detectar que este mal uso está asociado a otras variables, como la edad.

En los procesos de DM se suelen distinguir varias fases: en primer lugar, se realiza una exploración de la información disponible, y se extrae de las bases de datos corporativas de una empresa aquella información que se va a investigar, formándose una base de datos específica; posteriormente se formulan modelos o se definen pautas de comportamiento de interés; finalmente se procede a una validación de los resultados aplicándolos a nuevos conjuntos de datos.

El interés despertado en el ámbito empresarial ha sido tan grande que se están impulsando nuevos desarrollos teóricos para analizar grandes volúmenes de datos, utilizando una combinación de técnicas estadísticas y de inteligencia artificial, unido a la investigación en estructuras y bases de datos.

El éxito en la aplicación de las técnicas de DM depende, en primer lugar, del conocimiento de los datos disponibles, de la formulación precisa del problema que se trata de resolver, y en la utilización de las fuentes adecuadas de datos, tanto internos como externos.

Finalmente hay que hacer alusión a un problema que se presentará cada vez con más frecuencia en los procesos de DM: la confidencialidad y privacidad de la información. Por ejemplo, al hacer una transacción con una tarjeta de compra, la recogida de información está destinada a contabilizar esta operación. Se plantean preguntas sobre la licitud de emplear estos datos, que además no son anóminos, para otros fines, como los expuestos anteriormente: perfiles de clientes, ventas cruzadas, u otros estudios, que pueden atentar contra la intimidad de la persona, pues no es con este fin para el que ha proporcionado la información. En todo caso, a la hora de recoger microdatos, es imprescindible indicar para que se van a utilizar esta información, tanto directamente como en posibles estudios y minería de datos, el cual podría incluso ser utilizado por consultores externos, si no se especifica con claridad a los suministradores de información, para que fines puede ésta ser utilizada, y que limitaciones desean imponer a estos usos.

ALGUNAS HERRAMIENTAS INFORMÁTICAS

Los programas de DM han sido desarrollados en su mayor parte por las empresas que construyen paquetes estadísticos, aunque existe una oferta muy amplia en el mercado, como se puede comprobar con la lista de webs que aparece al final del texto.

Clementine es un software especializado en minería de datos al que puede acceder cualquier usuario. La particularidad que ofrece esta herramienta es la sencillez de su

manejo puesto que construye un sistema analítico que permite visualizar el proceso de negocio con los datos que le ofrece la base de datos que se desea estudiar. Entre las características de manejo se destacan las siguientes: Acceso a los datos: (Importa directamente de las bases de datos más importantes como Oracle, Ingres, Sybase e Informix, permitiendo importar datos de cualquier otra base o fichero con el comando Open file format), el filtrado simple y personalizado de registros, crear y renombrar registros (incluye funciones para procesar secuencias de éstos; es ideal para series de tiempo). La visualización de datos permite gran variedad de formatos, diagramas de puntos, histogramas, señalando zonas de interés. Incorpora redes neuronales y la inducción que son utilizadas para automatizar el proceso de toma de decisiones. A través de estas técnicas, *Clementine* aprende de salidas previas a realizar predicciones y juzgar nuevos casos. Con la programación visual el usuario no necesita conocer un programa complejo pues a través de iconos, ediciones, conexiones... se indica cómo leer y manejar los datos. Para la introducción, se utilizan modelos de árboles de decisión. También se incluyen algoritmos que encuentran normas de asociación en los datos. Utilizan K-medidas para la segmentación de los datos en grupos significativos y la red de Kohonen como algoritmo de segmentación. Permite combinar varios modelos en la predicción, lo que suele ofrecer mayor poder predictivo que cada uno de los modelos por separado.

Marksman es otro sistema de análisis de bases de datos orientado al marketing empresarial. Basado, principalmente en redes neuronales permite al igual que *Clementine* el análisis, clasificación y modelización de los datos ofreciendo su segmentación en grupos de interés para la empresa y hallando patrones de comportamiento y tendencia de los mismos. La mayoría de las características del programa son similares a *Clementine*. No obstante, ofrece una ventaja sobre el anterior dado que posee herramientas de mayor potencia a la hora de crear informes con los resultados obtenidos con las bases de datos analizados. Por su parte, *Clementine* es más idóneo para combinar con programas como SPSS que posee mayor variedad a la hora de crear estos informes y lo complementa en este aspecto. Sin embargo, la sencillez de visualización de *Clementine*, se hace menos patente en *Marksman*, que normalmente requiere de mayor tiempo de aprendizaje puesto que su entorno, aunque intuitivo, lo es en menor medida.

Knwlegde Acces Suite ofrece un paso más dentro de Data Mining. Crea una intranet con bases de modelos hallados sobre los datos y a la que los usuarios pueden acceder fácilmente. Utiliza el lenguaje PQL (lenguaje de consulta a modelos) diseñado para el descubrimiento de información y de estructura similar a SQL. El sistema de transferencia de conocimiento permite al usuario interconectar con el sistema web que transfiere el conocimiento y permite al usuario interconectar con el sistema web la información que se requiere conocer por propia inciativa del sistema. Posee ciertas características ventajosas que amplían oferta de las herramientas exclusivas de DM tales como mayor rapidez y eficacia (no volviendo a modelizar los datos por cada usuario), mejor precisión (al no tomar muestras, sino modelizar sobre la base de datos completa) etc.

Bibliografía

Adriaans, Pieter; Zantinge Dolf (1998). *Data Mining*. Addison Wesley Pub. Co.

Berry, Michael J. A.; Gordon Linoff (2000). *Mastering Data Minig*. John Wiley

Cabena, Peter; Pablo Hadjnian; Rolf Stadler; Jaap Verhess, Alessandro Zanasi (1997). *Discovering Data Mining from Concept to Implementation*. Prentice Hall.

Caridad y Ocerin, José M. (1998). *Econometría: modelos econométricos y series*

temporales. Reverté. Barcelona.

Groth, Robert (1999). *Data Mining: Building Competitive Advantage*. Prentice Hall.

Han, Jiawei; Micheline Kamber (2000). *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers.

Hand D. J. (2000). *Compstat 2000*. Jelke G. Bethlehem, Peter G. M. van der Heiden (ed.) Physica-Verlag. Heidelberg-New York.

Pyle, Doriam (1999). *Data preparation for Data Mining*. Morgan Kaufmann.

Westpjhal, Chris; Teresa Blaxton (1998). *Data Mining Solutions: Methods and Tools for Solving Real-World Problems*. John Wiley and Sons.

Data Mining and Knowledge Discovery. Journal, Kluwer Academic Publishers.

Data Mining (1999). SAS Institute.

DM Algorithms. (1999). Megaputer.

Webs asociadas a DW y DM

dw.ittoolbox.com
www.accrue.com
www. asacorp.com
www.blogicsys.com
www.catalysttech.com
www.cs.wits.ac.za
www.data-mines.com
www.datamining.com.sg
www.dwinfocenter.org
www.get-answer.com
www.isoft.fr
www.megaputer.ru
www.norsys.com
www.patternwarehouse.com
www.sgi.com
www.themeasurementgroup.com
www.trajecta.com
www.ultragen.com
www.abtech.com
www.almaden.ibm.com/cs/quest
www.bera.com
www.bluedatainc.com
www.cs.bham.ac.uk
www.data-mine.com
www.data-miners.com
www.digimine.com
www.econometria.com
www.hncmarksman.com
www.kdnuggets.com
www.neuralsystems.com
www.opin.com
www.schenley.com
www.spss.com/datamine
www.tnuinet.com
www.twocrows.com