ISBN 84-95336-48-0

COTEC es una fundación de origen empresarial que tiene como misión contribuir al desarrollo del país mediante el fomento de la innovación tecnológica en la empresa y en la sociedad españolas.

Cotec

Fundación Cotec
para la Innovación Tecnológica
Pza. Marqués de Salamanca 11, 2º izda.
28006 Madrid
Telf. (34) 91 436 47 74
Fax. (34) 91 431 12 39
http://www.cotec.es

ADE (CASTILLA Y LEÓN)
ADER (LA RIOJA)
AGENCIA NAVARRA DE INNOVACIÓN Y TECNOLOGÍA
ALSTOM ESPAÑA
ASOCIACIÓN INNOVALIA
AYUNTAMIENTO DE GIJÓN
AYUNTAMIENTO DE VALENCIA
BILBAO BIZKAIA KUTXA
CAJA DE AHORROS Y MONTE DE PIEDAD DE MADRID
CAJA DE AHORROS Y PENSIONES DE BARCELONA
CÁMARA DE COMERCIO E INDUSTRIA DE MADRID
CLARKE, MODET & Co
CONSEJERÍA DE CIENCIA Y TECNOLOGÍA (CASTILLA-LA MANCHA)
CONSEJERÍA DE INNOVACIÓN, CIENCIA Y EMPRESA (JUNTA DE ANDALUCÍA)
CONSULTRANS
DELOITTE
DIRECCIÓN GENERAL DE INVESTIGACIÓN DE LA COMUNIDAD DE MADRID
DIRECCIÓN GENERAL DE INVESTIGACIÓN Y DESARROLLO (GALICIA)
DMR CONSULTING
EADS ASTRIUM-CRISA
ELIOP
ENCOPIM
ENDESA
ENRESA
ERICSSON
EUROCONTROL
EUSKALTEL
FREIXENET
FUNDACIÓN AUNA
FUNDACIÓN BANCO BILBAO VIZCAYA ARGENTARIA
FUNDACIÓN BARRIÉ DE LA MAZA
FUNDACIÓN CAMPOLLANO
FUNDACIÓ CATALANA PER A LA RECERCA
FUNDACIÓN FOCUS-ABENGOA
FUNDACIÓN IBIT
FUNDACIÓN LILLY
FUNDACIÓN RAMÓN ARECES
FUNDACIÓN UNIVERSIDAD-EMPRESA
FUNDACIÓN VODAFONE
FUNDECYT (EXTREMADURA)
GRUPO ACS
GRUPO ANTOLÍN IRAUSA
GRUPO DURO FELGUERA
GRUPO LECHE PASCUAL
GRUPO MRS
GRUPO PRISA
GRUPO SPRI
HIDROELÉCTRICA DEL CANTÁBRICO
HISPASAT
IBERDROLA
IBERIA
IBM
IMADE
IMPIVA
INDRA
INSTITUTO DE FOMENTO DE LA REGIÓN DE MURCIA
INSTITUTO DE DESARROLLO ECONÓMICO DEL PRINCIPADO DE ASTURIAS
INSTITUTO TECNOLÓGICO DE ARAGÓN
INTEL CORPORATION IBERIA
MERCAMADRID
MERCAPITAL
MIER COMUNICACIONES
NECSO
OHL
O-KYAKU
PATENTES TALGO
PROEXCA
REPSOL YPF
SANTANDER CENTRAL HISPANO
SEPES
SIDSA
SIEBEL SYSTEMS ESPAÑA
SOCINTEC
SODERCAN (CANTABRIA)
TECNALIA
TÉCNICAS REUNIDAS
TELEFÓNICA
UNIÓN FENOSA
ZELTIA

FUNDACIÓN COTEC PARA LA INNOVACIÓN TECNOLÓGICA

DOCUMENTOS COTEC SOBRE OPORTUNIDADES TECNOLÓGICAS

21

MINERÍA DE DATOS

DOCUMENTOS COTEC SOBRE OPORTUNIDADES TECNOLÓGICAS

Primera edición:
Noviembre 2004

Depósito legal: M. 48.518-2004
ISBN: 84-95336-48-0

Imprime:
Gráficas Arias Montano, S.A.

ÍNDICE

1 PRESENTACIÓN

La Fundación Cotec para la innovación tecnológica tiene como una de sus actividades permanentes, desde hace más de doce años, resaltar aquellas oportunidades que permitan al tejido empresarial incrementar su desarrollo tecnológico, su capacidad de innovación y su competitividad.

Los Documentos Cotec sobre Oportunidades Tecnológicas constituyen una colección orientada al cumplimiento del objetivo estratégico de actuar como motor de sensibilización a la actitud innovadora en la empresa. Estos documentos se editan después de un proceso de debate para identificar los retos y oportunidades que ofrecen esas tecnologías y su aplicación por las empresas, así como los beneficios que les reportan.

Para el debate, la Fundación Cotec reúne a un cualificado grupo de expertos empresariales, de investigadores del sistema público de I+D y de consultores especializados, para que analicen las posibilidades de aplicación de esas tecnologías o servicios, y las oportunidades que ofrecen para los distintos sectores empresariales. En esta ocasión, se presenta el resultado de la sesión dedicada a la **Minería de datos,** que tuvo lugar en Madrid el día 24 de junio de 2004, en la sede de Cotec.

El objetivo principal de este documento es reflexionar sobre las técnicas y las aplicaciones de la minería de datos en España y sensibilizar a las empresas sobre los beneficios potenciales de emplearlas. El documento se centra en los fundamentos de esta tecnología, sus funciones básicas y el proceso para su utilización, describiendo posteriormente una serie de aplicaciones reales muy interesantes para terminar con una revisión del estado actual de la misma.

El equipo de expertos que participó en la sesión fue coordinado por los profesores Aníbal Figueiras y Ángel Navia de la Universidad Carlos III de Madrid, quienes, a su vez, prepararon el material de esta publicación. Cotec quiere dejar constancia de su agradecimiento a todos ellos, sin cuyo trabajo, comentarios y sugerencias este documento no hubiera sido posible.

Fundación Cotec.

SESIÓN COTEC SOBRE MINERÍA DE DATOS

Expertos coordinadores

- Aníbal Figueiras
  Universidad Carlos III de Madrid
- Ángel Navia
  Universidad Carlos III de Madrid

Expertos participantes

- Christian Blaschke
  ALMA Bioinformática
- José Dorronsoro
  Instituto de Ingeniería del Conocimiento
- Francisco Freire
  El Corte Inglés
- Carlos Gil
  SAS Institute, S.A.
- Luis Méndez del Río
  SAS Institute, S.A.
- Jorge Navas
  Ministerio del Interior
- Pablo Redondo
  Telefónica Móviles España
- Jorge Rubio
  SEGITUR, S.A.
- Alfonso Ventura
  Indra

2

EL VALOR OCULTO EN LOS DATOS

2.1. EL EFECTO LIBRO DE ARENA

Según estudios de la Universidad de California en Berkeley, la actual producción mundial de nueva información es del orden de las decenas de exabits (10^{18} bits) por año. El justamente prestigioso investigador italiano Roberto Saracco cifraba hace tres años la capacidad de transmisión de las redes de telecomunicaciones en el orden de los petabits (10^{15} bits) por segundo, duplicándose anualmente.

Tan insólitas cantidades dirán muy poco a la mayoría de las personas: los humanos no estamos bien preparados para interpretar magnitudes inusuales. Mucho más ilustrativo resultará sin duda hacer ver que un exabit equivale aproximadamente al contenido de un tomo de la Enciclopedia Británica por cada tres habitantes de la tierra, y que un petabit por segundo permitiría la transmisión de más de 300 millones de canales de televisión.

Pero no cabe asombrarse por tan desconcertantes equivalencias: la producción incluye, por ejemplo, todos los vídeos domésticos, y por las redes viajan todos los intercambios de archivos digitales. Hay muchos datos: pero la información (útil) y el consiguiente conocimiento (efectivo) no alcanzan tales volúmenes ni crecen a semejantes tasas. Por ello, autores hay que bautizan el efecto de estos exuberantes números con el difícil neologismo «infoxicación»,

induciendo a concluir que existe un dañino exceso de información, con lo que se puede llegar a despreciarla. Mucho más afortunado parece el símil de la Biblioteca de Babel, título de un conocido cuento de Jorge Luis Borges: en un cierto espacio compuesto de innumerables habitaciones exagonales se almacenan libros que contienen todas las combinaciones posibles de los símbolos de la escritura. Y todavía mejor el de *El Libro de Arena,* otra narración corta del genial argentino: habla de un libro que, como el cuerpo euclídeo de planos, consta de infinitas hojas. Preferible es esta segunda analogía porque, como hoy ocurre, pone tal almacén al inmediato alcance de los dedos. Las narraciones demuestran no sólo la inutilidad, sino también el real peligro de tamaños depósitos de escritura. Como siempre, Borges añade a una forma original e impecable una visión profundamente inteligente: no es lo mismo información que símbolos o datos. Sólo se obtiene valor cuando se puede —se sabe— manejar los datos.

2.2. DATOS, INFORMACIÓN Y CONOCIMIENTO

De poco ha servido hasta hoy que otro genio del pasado siglo, Claude E. Shannon, corto tiempo después de haber presentado una definición matemática de información, advirtiese en el breve artículo «The Bandwagon» de que otros campos del saber no podían valerse de una formulación puramente cuantitativa, para él necesaria porque su problema era ingenieril: determinar cuánta información podía transmitirse a través de un canal de comunicación dado, sin deteriorarla. De nada vale aún hoy la presencia de la milenaria raíz «forma» en el término información: demostrativa por vía etimológica de que hace muchísimo tiempo que se sabe que nuestros sentidos (y sistemas perceptivos) precisan de estímulos del tipo adecuado. Escasamente útil ha sido que, imperfecta y tal vez inconsciente-

mente, el márquetin de las compañías operadoras de Telecomunicaciones haya desenterrado las raíces del problema: cualquier cosa, en cualquier momento, en cualquier lugar... Un modo de reconocer que en la relevancia (adecuación, oportunidad, etc.) de la información para quien la obtiene o recibe está lo decisivo.

Los datos no son información porque no tienen el aspecto adecuado para su asimilación por quien los recibe. E incluso una máquina que recibe datos para guiar su funcionamiento no podría aprovecharlos si fuesen cualesquiera y no convenientemente estructurados. Claro que los datos contienen información: en muchos casos valiosa, ocasionalmente hasta insospechada...; pero nosotros, los humanos, no podemos desentrañarla sin las herramientas apropiadas. De ello hay cuantiosísimas pruebas: si el lector busca la ley de formación de la figura 1, empleada en test perceptivos por la Universidad de Michigan, le resultará bastante difícil; pero la dificultad decrece si se molesta en colorearla con distinto color para las casillas de cada figura geométrica (ver solución en el apéndice III).

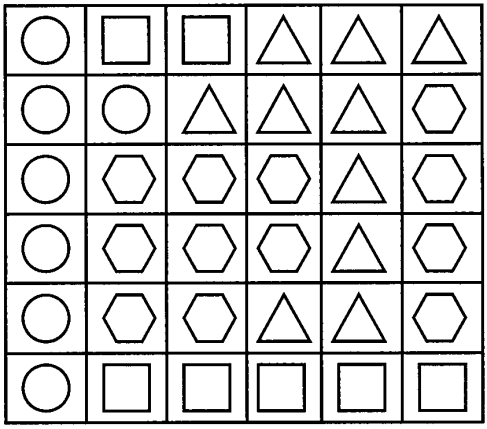

Figura 1

Test perceptivo de la Universidad de Michigan. La sustitución de las figuras geométricas por dígitos o, mejor aún, por colores, facilita grandemente la percepción de la estructura espiral subyacente (solución disponible en el apéndice III).

Y sin información no hay conocimiento. Todo lo que sabemos procede, directamente o tras elaboración, de lo que percibimos a través de nuestros sentidos, en un proceso de internalización de la información que, siguiendo a Jean Piaget, consiste en una fase de aprehensión seguida de otra de asimilación o estructuración: de puesta en contexto con nuestros otros conocimientos y con nuestras emociones (también nacidas del conocimiento), adquiriendo una organización radicalmente diferente de las computacionales —de ahí que aparezca el corazón («cor, cordis») en el vocablo «recordar»—, cuya mayor utilidad está acreditada por la sorprendente capacidad de los humanos para controlar el entorno.

Pese a tal evidencia, y al no despreciable crecimiento de la minería de datos («Data Mining») en los últimos años (en técnicas, en herramientas, en servicios y en aplicaciones), buena parte de quienes tienen acceso a datos importantes para su actividad —porque son datos de su actividad— y que les posibilitarían mejorarla ayudándoles a tomar decisiones que la orientasen según lo que realmente ocurre, no son conscientes de la ventaja que obtendrían del manejo de los instrumentos que le permitirían «ver» la información que dichos datos ocultan y llevar a cabo la subsiguiente interpretación. Y así en todos los ámbitos: una compañía con muchos millones de clientes necesitará de más potencia de cálculo que una con pocos miles; pero el beneficio relativo de emplear la minería de datos no tiene por qué ser distinto. Así, pues, ha de decirse que la minería de datos es potencialmente útil tanto para las grandes organizaciones como para las pequeñas y medianas empresas.

2.3. LA MINERÍA DE DATOS Y SU ENTORNO

Se llama minería de datos al proceso de extraer de una base de datos estructurada [1] la información relevante para los propósitos del agente y —segundo y crucial paso que a veces se olvida— analizarla para interpretarla: para darle sentido. Así se llega a adquirir conocimiento, que da ventaja para la actuación fundamental de las personas (u organizaciones): la toma de decisiones, para lo que, afortunadamente y por muy claras razones biológicas, sí estamos bien dotados... una vez que disponemos del debido conocimiento. Lo anterior puede representarse —tomándose ciertas libertades simplificadoras— según el diagrama de la figura 2, en el que queda claro el carácter instrumental de la minería de datos. Y ha de insistirse, de acuerdo con lo que antecede y con la propia figura, en que se trata de un proceso: de una actividad que implica recursos materiales y actuación de personas, y no simplemente de un sistema que procese datos.

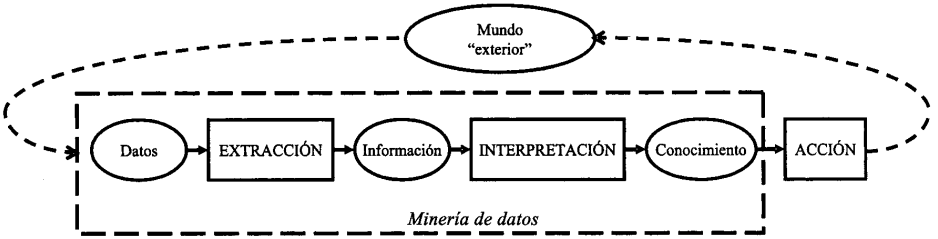

Figura 1

El papel de la minería de datos

Si bien en un pasado ya remoto, con el «mundo exterior» fraccionado en reducidas porciones poco interdependien-

[1] Trabajar a partir de los datos contenidos en un repositorio estructurado es lo típico a partir de mediados de la década de los ochenta: hasta entonces no era frecuente disponer los datos de tal modo.

tes entre sí, la minería de datos tenía menor razón de ser, la masificación y, sobre todo, la globalización (recuérdense las cifras con las que se ha abierto este capítulo) sugieren que hoy no es prescindible, y que mañana lo será menos (para los que sobrevivan, permítasenos advertir). Toda esperanza de un idílico retorno a la obtención cómoda y directa de información es vana: los datos numéricos y categóricos tienen su propia esencia y un rapidísimo crecimiento, y en otros casos incluso la información percibible por vía sensorial se digitaliza para lograr ventajas en computación y transmisión, incrementando el provecho que podemos obtener de las máquinas al reducir la probabilidad de errores (nada antinatural: también nos comunicamos construyendo un lenguaje a partir de símbolos, los fonemas).

Claro está que la «minería» no tiene por qué restringirse al manejo de datos propiamente dichos: la situación es equivalente cuando se dispone de información directamente asimilable, como la escrita, pero en volumen tal (o con tan numerosos objetivos) que resulta imposible examinarla para los humanos. De ahí la aparición de la minería de textos; y, de similar manera, la emergencia de la minería multimedia (audio, imagen fija, vídeo, gráficos, etc.), todavía padeciendo importantes cuellos de botella en extracción de rasgos (características) e indexado, pero de dorado futuro. Ambas técnicas de minería comparten similitudes y referencias con la minería de datos, pero aquí no cabe extenderse en la correspondiente discusión.

Como hemos dicho, se reserva el nombre de minería de datos para los procesos que incluyen extracción de información de datos localizados en un repositorio estructurado (general, «Data Warehouse», o específico, «Data Mart»). Cuando los datos se encuentran en entornos no estructurados, como en la WWW, se ha dado en llamar al proceso descubrimiento del conocimiento («Knowledge Discovery»); naturalmente asociado al caso multimedia. Muy preferible parece la denominación alternativa minería en

la red («Web Mining»), por analogía con la minería de datos clásica: el conocimiento no se descubre, se adquiere en el proceso. Lo que se ha bautizado como gestión del conocimiento («Knowledge Management») supone un abuso menor, ya que trata no sólo de aprovechar y estructurar la información disponible por y en una organización, sino el conocimiento de las personas que la integran mediante su explicitación o por transferencia tácita, e incluso propiciar la creación de conocimiento: involucrando así directamente a las personas (por lo que son de fundamental importancia los aspectos humanos en este proceso).

No es propósito de este trabajo encarar la discusión de todas estas disciplinas: excedería con mucho los límites previstos, cuando no las capacidades de los autores. Pero sí hacer ver que la minería de datos tiene fuertes relaciones con los otros procesos citados, y no está de más recordar que generalmente la mejor comprensión de un campo y los mayores avances en él provienen del examen de su contexto y convenientes extrapolaciones: casi todos los avances relevantes, desde los de las ciencias básicas hasta las aplicaciones tecnológicas, provienen de ensayar aproximaciones acreditadas en otras áreas y adecuadamente modificadas de acuerdo con un buen conocimiento del caso particular bajo análisis, que es lo que ocurre también en la minería de datos. Por ello resulta recomendable que quienes desempeñen tareas relacionadas con la minería de datos no dejen de interesarse por las otras minerías aquí presentadas.

3

LOS FUNDAMENTOS DE LA MINERÍA DE DATOS

3.1. LAS FUNCIONES BÁSICAS

Debe de resaltarse, como comienzo, que la minería de datos es un recurso para un objetivo de negocio, para cuya consecución sería importante conocer respuestas a preguntas que ni siquiera se sabe formular (en caso contrario, puede recurrirse a lo que se denomina procesado analítico «On Line» («On Line Analytical Processing», OLAP). Pero, aun así, resulta esclarecedor presentar el proceso como un conjunto de respuestas.

Los datos que se manejan en los procesos de minería, también conocidos como observaciones, instancias, muestras o ejemplos, son una colección de vectores de variables o rasgos; vectores en muchos casos etiquetados en su totalidad o en parte: es decir, incluyen la respuesta a una pregunta a la que el que los maneja desea contestación. Tales instancias (etiquetadas) se han obtenido como resultado de registrar lo acontecido en previas experiencias o experimentos, y son análogas a aquellas a las que se va a aplicar un diseño obtenido durante el proceso de minería: el de una función que hace corresponder a cada vector de variables observables una de las repuestas posibles, su (desconocida) etiqueta.[2]

[2] Esta formulación, conocida como predictiva, ya que sirve para dar respuestas ante situaciones en las que no se conocen, es la de mayor in-

Así, en el caso de concesión de un crédito o préstamo, los datos etiquetados corresponden a clientes que lo han solicitado en el pasado, incluyendo típicamente variables de tipo sociodemográfico (edad, situación familiar, antigüedad en el empleo, ingresos, etc.), de relación económica con el otorgante (saldo medio y antigüedad de cuentas, otros instrumentos financieros, etc.) y de la propia operación (objeto, importe, etc.), con la etiqueta de morosidad o no (obsérvese que hay un efecto de recorte: no hay etiqueta para los casos denegados).

Las preguntas cuya contestación se pretende (para los nuevos casos que se presenten) no difieren en nada de las habituales que una persona se hace continuamente: ¿sucede o va a suceder algo?; ¿qué?; ¿por acción de quiénes?; ¿a quiénes?; ¿cuánto?; ¿cuándo?... Por tanto, nos encontramos sistemáticamente con problemas de tipos básicos iguales: detección de un cierto suceso, como ocurre en control de fraude; clasificación de un caso, como en segmentación de mercados (que, eventualmente, puede convenir llevar a cabo mediante un previo agrupamiento, o «clustering», si no se dispone de etiquetas: para un subsiguiente etiquetado de los grupos a través de la exploración de algunos de sus representantes); decisión sobre qué opción elegir entre varias posibles (que, ocasionalmente, toma la muy particular forma de la asociación, como la de productos de venta conjunta más frecuente en el análisis de cestas de compra: un problema de decisión peculiar que se resuelve mediante métodos estadísticos elementales); estimación de una cierta magnitud, como el riesgo o el beneficio esperado de una operación, lo que recibe el nombre de predicción cuando se trata de valores futuros.

---

terés. La formulación exploratoria, en la que se manejan los datos para concluir si existen relaciones entre ellos, es también posible: pero, en cierto sentido, menos relevante. En esta segunda formulación pueden incluirse, por ejemplo, los procesos de agrupamiento o segmentación y los de establecimiento de relaciones y su análisis.

Tal complejidad aparente se reduce conceptualmente a sólo dos tipos de problemas: los de detección/clasificación/decisión, en los que hay que optar por una entre un conjunto discreto de alternativas (hay/no hay, clase A, B..., o Z, sí/no...), que suelen denominarse hipótesis, y los de estimación (predicción), en los que hay que atribuir un valor a una magnitud que no se puede observar directamente.

Pero antes de presentar la visión abstracta de estos problemas, y para facilitar al lector la identificación de casos reales con ellos, ilustraremos ambos tipos conceptuales mediante ejemplos sencillos.

La concesión de préstamos y créditos es una de las actividades tradicionales de muchas entidades financieras. Estas disponen de datos sobre operaciones pasadas similares de las que conocen el resultado (morosidad o no): la que se llama «etiqueta» de los ejemplos conocidos. Cada caso incluye una larga lista de variables de diversos tipos (características de la propia operación: importe, plazo, objeto, etc.; datos sociodemográficos sobre el solicitante: profesión y antigüedad en ella, ingresos anuales, propiedades...) y datos de tipo financiero procedentes de la propia entidad o de otras (cuentas y su antigüedad y saldos medios, tarjetas de crédito, historial de operaciones anteriores similares...). A partir de este tipo de datos ha de realizarse una decisión para los nuevos solicitantes: conceder o no el préstamo o crédito. Para disponer de una ayuda para la toma de esta decisión ha de recurrirse a extraer de la base de datos etiquetados de que se dispone la información necesaria para maximizar el beneficio del proceso (no necesariamente la probabilidad de acierto).

Un problema de la misma familia es el constituido por los procesos de detección de fraude: en operaciones con tarjetas de débito o de crédito, en la utilización de servicios de telefonía, etc. En estos casos se dispone también de datos etiquetados, correspondientes a casos de fraude y no fraudulentos, y que incluyen como variables también datos

sociodemográficos del usuario y el historial previo de utilización de la tarjeta o del servicio. De nuevo se trata, en lo básico, de un problema del primer tipo conceptual: detectar (y, la mayoría de las veces, a tiempo) la aparición de casos de fraude.

En esta categoría entran también muchos de los problemas de fidelización de clientes: desde la previsión de bajas de tomadores de una cierta clase de pólizas en el ámbito de los seguros, hasta los estudios de rotación («churn»: paso de los clientes de un proveedor a otro) en compañías de servicios, como los de telecomunicaciones. Dado que más adelante, en el capítulo 4 dedicado a aplicaciones de la minería de datos, se describe en detalle, entre otros, un ejemplo real (el programa «Customer First» de MCI), remitimos al lector a ese lugar para una primera visión de esta familia de problemas.

No menos abundantes son los problemas de estimación: aunque es bien cierto que, generalmente, se trata de procesos cuya utilidad se halla en una posterior toma de decisiones. Así, puede tratarse de predecir el volumen de ventas de un cierto producto a partir de datos de mercado pasados y de acciones esperadas de la competencia (normalmente para decidir qué acciones se han de tomar: en publicidad y márquetin, ventas, innovación, sustitución...); otro ejemplo, que se puede abordar con técnicas sencillas, es la estimación de cifras de ventas conjuntas de pares o grupos de productos: el ejemplo más conocido es el de la cesta de la compra en supermercados/hipermercados, que es posible mediante un simple conteo de pares en tiques de caja, y que en los Estados Unidos hizo conocer ya hace años el (en apariencia) sorprendente resultado de que pañales y latas de cerveza constituían un par de adquisición conjunta muy frecuente (sin que ello implique que lo mismo ocurra en otros lugares... ni en otros momentos). Este proceso de estimación puede utilizarse, por ejemplo, para orientar en la política de colocación de productos en las estanterías.

Que los procesos de decisión y estimación se realicen simultáneamente no es infrecuente: muchas veces se desea en una decisión estimar la probabilidad de acierto; por ejemplo, la probabilidad de morosidad en concesión de préstamos o créditos. En tal caso, se habla de decisión (o clasificación) puntuada o cualificada. También es posible superponer un umbral para tomar decisiones en problemas de estimación: como mantener o no un producto en el mercado según se supere o un cierto nivel de beneficios.

Las figuras 3 y 4 representan lo esencial de tales problemas: los datos de que se dispone están relacionados estadísticamente [3] con las hipótesis o el parámetro que se vayan a estimar y, a partir del conjunto de datos etiquetado de que se dispone, ha de diseñarse el decisor o el estimador: como se ha dicho, una función que hace corresponder al dato genérico $\underline{x}$ una decisión D_i o un valor estimado $\hat{s}$.

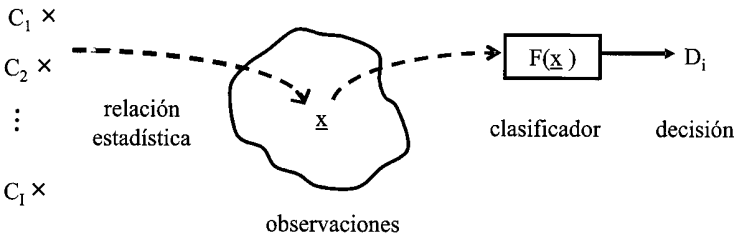

Figura 3

Visión estructural del problema de clasificación (decisión)

[3] En otros ámbitos, la relación viene dada por una transición probabilística que tiene lugar en el mundo físico, como cuando se observa una señal electromagnética perturbada por un ruido. Si es así y se conocen las leyes físicas correspondientes, pueden construirse soluciones analíticamente, recurriendo sobre todo a la teoría bayesiana. Aunque todo esto es totalmente inhabitual en minería de datos, donde hay un conocimiento escaso y muy difuso de lo que realmente está ocurriendo, nos permitimos recomendar a los profesionales que no lo olviden, ya que permite utilizar conceptos útiles y diseñar soluciones indirectas (mediante modelado) e híbridas.

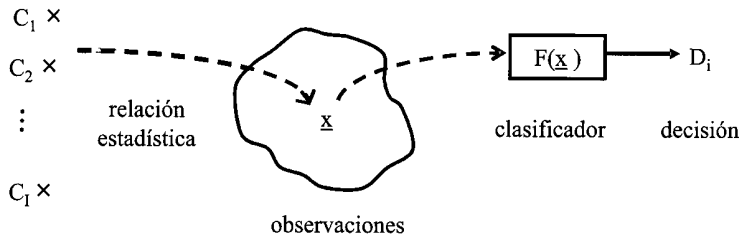

Figura 4

Visión estructural del problema de estimación

El caso del agrupamiento es análogo al de clasificación en el sentido de que su resultado es el establecimiento de grupos como lo es el de clases en clasificación: pero en agrupamiento se sustituye el papel de las etiquetas por la maximización de un adecuado, o al menos razonable, criterio de semejanza entre los datos que compongan cada grupo.

El agrupamiento, que en ciertos ámbitos se denomina segmentación, ofrece como resultado un conjunto de grupos de muestras semejantes (según el criterio empleado), y en muchas ocasiones con un representante o prototipo de cada grupo. Se emplea mayoritariamente en procesos exploratorios: es más fácil evaluar conjeturas cuando se trabaja con un número reducido de prototipos que cuando se ha de considerar una muy elevada cantidad de ejemplos. La representación por prototipos también permite llevar a cabo diseños preliminares orientativos para los procesos de clasificación y de estimación. Y la segmentación resulta muy ventajosa cuando, para obtener etiquetas, hace falta recurrir a la experimentación: así, si se quiere dirigir una campaña de márquetin por correo a una población muy numerosa, y no se tiene un criterio fiable de qué parte de esa población es objetivo (puede estar interesada) en la oferta, proceder a una razonable segmentación previa y llevar a cabo un test reducido, dirigiéndose a unos pocos

individuos de cada segmento y etiquetando después los segmentos según los resultados (con posibles revisiones de la segmentación), permite reducir notablemente el coste de la campaña sin que decrezca apreciablemente su alcance. Ya desde este momento inicial, una vez que se sabe cuáles son sus funciones básicas, resulta procedente discutir una primera serie de dificultades en la aplicación de la minería de datos: curiosamente, pese a su carácter no técnico, son la causa de un porcentaje muy alto de fracasos.

3.1.1. De donde no hay no se puede sacar

Los datos sólo pueden proporcionar la información que hay en ellos: ni un ápice más. De modo que han de considerarse esenciales, sobre todo por parte de las organizaciones que desean obtener ventaja de la minería de datos:

- la obtención de un conjunto de datos (etiquetados) para el diseño que sea representativo del problema que se desea resolver: es decir, en número suficiente; o incluso más que suficiente, en caso de duda;
- la cuidadosa selección de las variables que integran dichos datos o la extracción de las características o rasgos apropiados para los ejemplos, en caso de gozar de tal posibilidad (independientemente de la posterior selección que pueda incluirse en el propio proceso de minería).

Disponer de ejemplos suficientes y de calidad (lo que implica la extrema importancia de las tareas de «limpieza» de los datos disponibles en bruto: imputación de valores perdidos, eliminación de muestras fuera de margen, etc.) constituye una necesidad insoslayable para el éxito de cualquier aplicación de la minería de datos: a la vista de esa disponibilidad se puede decidir si merece la pena iniciar el proceso o si, por el contrario, lo que conviene es re-

colectar más datos. Para ello, y además de las características del problema (p. ej., frecuencia y dispersión del suceso a detectar), resulta de la máxima utilidad la consideración de precedentes (comparables): que también puede arrojar luz sobre otros aspectos del trabajo, como la elección de las técnicas algorítmicas para el tratamiento de datos, las herramientas de «software», o las plataformas computacionales («hardware») en que llevar a cabo el proceso.
Nótese que lo que precede no implica en absoluto que sea preciso disponer de una ingente cantidad de ejemplos para aplicar la minería de datos: basta un conjunto que represente fielmente el problema que se quiere considerar.

3.1.2. Sólo interesan las respuestas a lo que no se sabe

Expresado de otro modo: de nada vale que el sistema diseñado, y particularmente el decisor o estimador, replique perfectamente las etiquetas de los datos empleados para el diseño, pues para eso bastaría con una mera memorización. Lo verdaderamente importante es lo que se conoce como *generalización* (o poder predictivo): la capacidad de proporcionar buenas respuestas (según el criterio de bondad que se decida aplicar) para los casos que se desea clasificar en la aplicación verdaderamente dicha del sistema diseñado. Tal característica ha de merecer la mayor atención de los diseñadores si quieren que su trabajo aporte la pretendida utilidad para los usuarios y sus clientes.
El estudio de la generalización no resulta sencillo, ni muchísimo menos: depende de (además del problema) los datos, la estructura elegida para el decisor o el estimador, el criterio de calidad que se aplique y del modo de llevar a cabo el diseño. Hay numerosos procedimientos analíticos y heurísticos para conseguir una buena generaliza-

ción, pero su exposición excede del ámbito de estas páginas.
También aquí son de importancia los factores humanos... Recurriremos ahora a frases comunes que se refieren a dificultades habituales.

3.1.3. Cada uno a lo suyo

Es lo que, consciente o inadvertidamente, dicen algunos expertos en el asunto o aspecto del negocio en que se puede aplicar la minería de datos: postulan así que, como tales expertos, se bastan ellos solos para obtener la necesaria información de los datos. Tal posición no es más que la consecuencia de un temor comprensible pero infundado: que el sistema se convierta en un rival que pueda desplazarlo... cuando en realidad, como ya se ha dicho, estaría disponiendo de un instrumento que potenciaría su actividad. Con ello, se pierde la posibilidad de un mejor diseño contando con el conocimiento del experto: de modo que, si el sistema llega a aplicarse, el experto ha logrado... perjudicarse a sí mismo.
A veces, simplemente se trata de que el experto no quiere realizar el esfuerzo que representa conocer, y aplicar, pero sobre todo interpretar, los resultados del sistema de Minería. Posteriormente discutiremos en mayor detalle esta renuencia: pero es obvio, como en el caso anterior, que el experto no percibe como incentivo las ventajas que el uso del sistema le puede dar. Si bien en el pecado tendrá la penitencia, conviene que los más altos responsables de la organización primen directamente tal esfuerzo, para propiciar que la posterior emergencia de las ventajas acabe de convencer al experto.
Con lo anterior, se aprecia que el real sentido de «cada uno a lo suyo» es justamente el contrario del aparente a primera vista: el experto, a actuar como tal, y la minería de datos, a servirle de ayuda.

3.1.4. No hay que meterse en lo que no te importa

La anterior es la frase popular más ajustada que hemos encontrado para referirnos al rechazo inicial que se produce en un muy elevado porcentaje de los clientes de la organización que pretende aplicar la minería de datos cuando los datos son personales. Otra vez se trata de un miedo y de una falta de percepción del potencial beneficio.

El miedo es claro: se teme que los datos puedan emplearse para objetivos completamente distintos de los esperables de la situación en la que se proporcionan. Las legislaciones ya se han ocupado de ello: respetarlas más que escrupulosamente es el único camino para el beneficio de todos.

Beneficio que muchas veces el cliente cuyos datos se manejan no alcanza a prever: hacerle comprender que existe y demostrárselo tan rápida y sólidamente como se pueda sirve para ganarse lo que resulta esencial, que es no sólo su anuencia, sino su activa cooperación a lo largo del tiempo.

Y, finalmente, en lo que se refiere a los técnicos:

3.1.5. Siempre se rompe la cuerda por lo más flojo

Los diseñadores no avezados suelen incurrir en el grave error de dedicar todo su esfuerzo al elemento aparentemente nuclear de la minería de datos —el clasificador o el estimador—, olvidando con ello dos cuestiones fundamentales.

La primera es que el experto ha de poder valerse del sistema para tomar sus decisiones o hacer sus recomendaciones: y para ello necesita interpretar los resultados. Facilitárselo con diseños tan sencillos como sea posible, e inclusive con ayudas específicas, tiene también una impor-

tancia crucial (lo que no implica liberar al experto de todo esfuerzo mediante diseños inapropiadamente sencillos: del punto de equilibrio hablaremos más adelante).
Segunda cosa olvidada: la minería de datos es un proceso; es decir, en lo esencial, es una sucesión de pasos. Por ello, la calidad del resultado queda limitada por la calidad del peor paso… y pasos hay muchos; de modo que hay que cuidar no los más delicados, sino todos ellos.
En virtud de esta razón, se opta aquí por invertir el habitual orden de presentación de los documentos que versan sobre la minería de datos: expondremos primeramente el proceso global de minería, incluyendo ciertas observaciones y advertencias sobre el mismo, para pasar después a las técnicas (que hacen referencia al núcleo computacional del proceso, el clasificador o estimador), de modo que la sombra de estas técnicas no obscurezca los demás aspectos.

3.2. EL PROCESO DE LA MINERÍA DE DATOS

La figura 5 presenta un esquema (simplificado) de los pasos que conforman un proceso aplicativo de minería de datos, mediante el que se busca información relevante para un aspecto del negocio claramente definido. En esta figura nos apoyaremos para la explicación de dicho proceso, pero desde el principio advertiremos que, típicamente, se hace preciso recorrer bucles varias veces: es decir, volver atrás y repetir una parte del camino. La necesidad de hacerlo sólo se pone abiertamente de manifiesto en las etapas en las que se intenta comprender o se comprende lo que se está obteniendo: es decir, las etapas que hemos denominado «conjeturas», donde a la vista de resultados (provisionales) del diseño y su análisis se pueden aventurar hipótesis sobre el problema que se está tratando, y «conocimiento», donde los resultados y el análisis de la efectiva aplicación del diseño arrojarán nueva luz sobre el proble-

ma. Por eso marcamos en la figura con flechas hacia atrás dichas etapas: de lo que en ellas se interprete se deducirán modificaciones potencialmente ventajosas de lo ejecutado en etapas anteriores, marcadas en la figura con flechas de llegada o entrantes, y así se configurarán bucles razonables y provechosos, evitando caer en un caos de avances y retrocesos. Cabe añadir que es regla sensata actuar, al trazar un bucle, primero sobre las etapas más cercanas al origen, si lo requieren. Esta explicación es fundamentalmente conceptual y, como tal, reducida: existen propuestas más detalladas de estandarización de metodologías de implantación de sistemas de minería de datos, como la denominada CRISP-DM [4] que, fruto de un proyecto «ESPRIT» financiado por la Comunidad Europea, aglutina los procesos de descubrimiento del conocimiento habitualmente usados en la industria y los actualiza en función de las necesidades de los usuarios finales hasta conseguir definir y validar una metodología a seguir en cualquier aplicación comercial de la minería de datos.

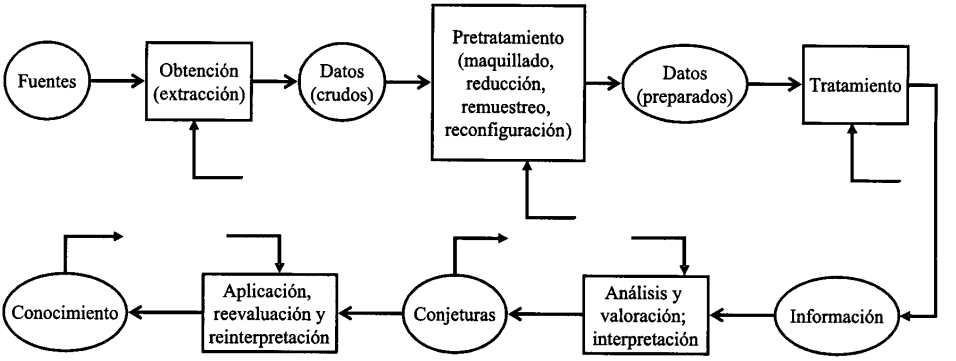

Figura 5

El proceso de minería de datos

[4] Cross Industry Standard Process for Data Mining, http://www.crisp-dw.org.

3.2.1. Obtención de datos (crudos)

De la importancia clave de los datos y de su calidad ya se ha hecho mención («De donde no hay…»): consecuentemente, ha de cuidarse su obtención (o la extracción de rasgos); lo que requiere una buena política de acceso a las correspondientes fuentes: que pueden incluir variables diferentes o ser heterogéneas en otros muchos sentidos (p. ej., en cuanto a su tipo: numéricas, categóricas, semánticas...), pero han de merecer un justificado crédito. El acceso directo mediante telecomunicaciones a al menos algunas de estas fuentes puede resultar crítica para ciertas aplicaciones, como por ejemplo la predicción en operaciones en tiempo real.
Los datos crudos han de almacenarse: el hecho de que se sometan a un pretratamiento para buscar una mejora de las prestaciones del proceso de minería no debe llevar a su destrucción, porque los datos pretratados pueden haberlo sido de muy diversas formas, pocas veces invertibles, y en cualquier momento es posible que se evidencie la conveniencia de sustituirlos por los resultantes de otro pretratamiento. Por ejemplo: puede ensayarse el empleo del cociente entre dos variables en lugar de éstas, pero esas dos variables han de conservarse para posibilitar que, si se sospecha que sería mejor elevar el numerador al cuadrado, pueda hacerse. Obviamente, recurrir a un buen sistema de almacenamiento («Data Warehouse») facilita el aseguramiento de la calidad de los datos y su procesamiento ordinario.

3.3.2. Pretratamiento

Confiar en que el adecuado diseño de un clasificador o estimador que actúe sobre los datos crudos va a proporcionar resultados óptimos, buenos, o al menos satisfactorios en el proceso de minería supondría aceptar, de entrada, un absurdo principio que ya se ha criticado con ante-

rioridad: que se presenten como se presenten las cosas, con suficiente «potencia de cómputo» será posible extraer su contenido informacional. Tal vez cabría pensar así si se dispusiese de prácticamente ilimitadas cantidades de ejemplos: pero eso no ocurre casi nunca, y, en todo caso, implicaría un también ilimitado crecimiento de recursos computacionales. De modo que no es habitual proceder así, sino ensayar transformaciones que la experiencia dice que pueden resultar convenientes. Además, los expertos suelen saber (al menos cualitativamente) en qué forma intervienen al menos algunas de las variables.

Así, pues, ha de considerarse que un pretratamiento de los datos antes de presentarlos al clasificador o estimador constituye un paso importante en la minería de datos. De hecho, cabe decir que una máquina elemental (p. ej., lineal) sería suficiente para resolver cualquier problema si los datos crudos se transformasen adecuadamente, ya que es posible demostrar que así se puede construir una máquina (en el sentido de algoritmo) aproximadora universal. Hay dos circunstancias que hacen del pretratamiento una etapa delicada en la práctica:

- no existen indicaciones *a priori* de cómo proceder (salvo las provistas por el conocimiento experto);
- cualquier desafortunada destrucción de información en el preprocesado de los datos es irreparable en el resto de proceso.

De modo que hay que comportarse con extrema cautela, procurando no eliminar ningún contenido que no pueda considerarse con certeza como «ruido» (es decir, como irrelevante para el problema que hay que resolver).

Aparte de recurrir al conocimiento de los expertos, existe un catálogo de operaciones que pueden ser adecuadas en muchas ocasiones, entre las que cabe destacar:

- Las que constituyen lo que se conoce como *limpieza* o *maquillado* de los datos (¡no de los resultados!); como

pueden ser la eliminación de valores manifiestamente erróneos o fuera de márgenes, la imputación de valores ausentes en algunos registros, el etiquetado de muestras sin etiqueta, etc.

- Las que componen los métodos de *reducción:* eliminación de variables, crudas o transformadas, irrelevantes o redundantes, o de muestras sin influencia sobre los resultados (muestras «no críticas»).
- Las técnicas de *remuestreo,* o creación de nuevas poblaciones, de las que hay muchos casos particulares con diversos propósitos: así, puede procederse a enfatizar —aumentar la frecuencia de aparición— muestras de carácter crítico —de clasificación poco clara— para que tengan mayor peso en el diseño, y mejorar la separabilidad de las diversas clases.
- Las técnicas de *reconfiguración* de los datos: transformando las muestras lineal o no linealmente, o aplicando codificaciones con el fin de resaltar la información más relevante para la resolución del problema.

Con todas estas operaciones pueden asociarse procedimientos analíticos (cálculo de estadísticas, relevancias, etc.), además de heurísticos.

Descender a un mayor detalle en esta taxonomía requeriría cientos de páginas: a falta de ellas, hemos de decir que hay muchas fuentes no despreciables de técnicas de pretratamiento en estadística clásica, tratamiento de señales, etc. Su conocimiento y utilización no está de más, aunque su adecuada elección, que muchas veces abre el camino al éxito, es más una cuestión de arte que de técnica analítica.

Debe aclararse, por último, que es casi siempre innecesario recurrir a la totalidad de los datos disponibles para llevar a cabo un proceso aplicativo de minería de datos: basta con una conveniente muestra para el diseño (separable en submuestras de entrenamiento y de validación) y otra para evaluación o test. Bien es cierto que habrá que

vigilar el funcionamiento para muestras que se obtengan tras haber implementado el diseño.

3.3.3. **Tratamiento** *(propiamente dicho)*

Se llama así a la aplicación del decisor o del estimador que se diseñe sobre los datos preparados por el pretratamiento; obteniendo una salida por cada nueva instancia que se presente. Obviamente, se trata de un paso clave en el proceso de minería, por lo que nos detendremos un poco en él.

Como ya se ha dicho, en las situaciones típicas de minería de datos no hay o no se dispone de un modelo físico subyacente, por lo que no cabe aplicar directamente técnicas analíticas para el diseño. Sí es posible recurrir a las técnicas semianalíticas, en las que, a partir de datos etiquetados (datos para los que se conoce, por haberla observado, la decisión o el valor a estimar correspondiente), se estima la información estadística precisa para aplicar una técnica analítica: probabilidades o densidades de probabilidad. Estas técnicas semianalíticas son delicadas, por requerirse como proceso adicional la estimación de información estadística; y, en todo caso, subóptimas, ya que dicha estimación se realiza de un modo que no va dirigido a obtener una buena solución del problema bajo análisis. Las técnicas o «métodos máquina», que se presentan a continuación, solventan el segundo inconveniente, ya que todo su diseño se encamina a obtener una buena solución, y, en muchas ocasiones, reducen la importancia del primero. Los métodos máquina son, en realidad, algoritmos que transforman un dato, un vector $\underline{x}$ cuyos elementos son los valores de las variables que corresponden a un cierto caso, en una salida o que aproxima la decisión que se ha de tomar o el valor que se debe estimar, y que se obtiene aplicando una función F (de valores discretos, para decisión) o f (de valores continuos, para estimación), función

que depende de un conjunto de parámetros o pesos $\underline{w}$: así, p.ej., F puede ser el signo de una combinación lineal de las variables de x, o lo que es lo mismo
$F = sgn(w_0+w_1x_1+\ldots+w_nx_N)$.

Los parámetros de la máquina, $\underline{w}$, se determinan aplicando los datos etiquetados a su entrada y forzar, buscando valores para ellos, que las correspondientes salidas se parezcan a las etiquetas, lo que se consigue midiendo el parecido mediante una adecuada función de coste, de papel análogo a las empleadas en los métodos analíticos. Hay que insistir en que el objetivo de este proceso no es conseguir el mayor parecido posible, sino que la máquina «aprenda» o «quede entrenada» para funcionar con buena generalización: respuestas satisfactorias ante los nuevos datos que se pretende clasificar en la aplicación práctica del diseño. Los procedimientos de búsqueda empleados pueden consultarse en infinidad de textos dedicados al tema, o bien de decisión y estimación, e incluso de minería de datos: no podemos detenernos aquí en su presentación y discusión.
Hay muchos tipos de máquinas: como quiera que revisarlos en la parte principal de este texto dificultaría su lectura, pero también es verdad que conocer sus características fundamentales resulta importante para muchos interesados, hemos decidido dedicar un anexo a una somera presentación de los principales, seguida de una exposición de sus ventajas e inconvenientes, y una brevísima indicación de las tendencias que consideramos más prometedoras. Por completitud, también se incluyen las ventajas e inconvenientes de algunos otros métodos: los ya mencionados semianalíticos (estadísticos) y los de visualización y preguntas. Por ello, el apéndice (I) se titula: «Las tecnologías para (el tratamiento en) la minería de datos». Ni que decir tiene que la selección de la(s) tecnologías(s) que se deben emplear y un apropiado diseño de la(s) máquina(s) constituyen aspectos críticos de la fase de tratamiento.

3.3.4. Interpretación

Cuando se ha diseñado un clasificador o un estimador, puede verificarse cómo actúa sobre las muestras de prueba. Es el (primer) momento en que el experto (preferiblemente acompañado por el diseñador, de no ser el mismo), además de valorar prestaciones, ha de buscar una interpretación del funcionamiento de dicho clasificador o estimador: ¿por qué da ciertos resultados ante ciertas muestras?; ¿qué es lo importante de los datos que se manejan?; ¿cómo puede expresarse la actuación en términos (fácilmente) comprensibles para un humano?; etc.

Obviamente, la interpretación es inmediata si se ha aplicado un sistema experto, es decir, el constituido por reglas del tipo «Si (A, B...) entonces (X, Y...)», explicitadas por expertos humanos o construidas a partir de los datos. Y tampoco es difícil para otros métodos máquina (vid. apéndice I)... hasta llegar a las redes neuronales: se tiende a pensar en ellas como «cajas negras», completamente inescrutables, razón por la que se rechaza en muchas ocasiones su empleo, pese a su acreditada superioridad potencial para la resolución de problemas claramente no lineales.

Dada la aparente dicotomía, han de dedicarse aquí dos palabras a lo que ha de llamarse interpretación. En el caso de los sistemas expertos, por ejemplo, el experto humano acepta la estructura del sistema para la interpretación... lo que no deja de ser curioso: el humano no interpreta los resultados, sino que se deja convencer por la arquitectura de la máquina... que podría ser (poco o mucho) inadecuada. Desde luego, en una situación forense, es cómodo repetir lo que la máquina dice con tanta contundencia, pero alegar que se ha interpretado el proceso es inapropiado, y tomar como razón algo cuya exactitud puede dejar mucho que desear puede, sin vacilación, calificarse de abusivo. Mucho más dificultoso es, desde luego, interpretar cómo procede una red neuronal; pero, si se logra, es una verdadera interpretación.

El análisis y la interpretación de los resultados del tratamiento de los datos permite establecer una primera serie de conjeturas sobre la importancia de las diversas variables, la conveniencia de determinadas transformaciones, el grado de éxito de los procedimientos de imputación utilizados, lo afortunado de la elección de un cierto modelo o arquitectura, lo acertado del criterio impuesto para la optimización y de la búsqueda aplicada..., incluso sobre lo satisfactorio del resultado general y las posibilidades de mejorarlo. En virtud de tales consideraciones conviene recorrer de nuevo, atentamente, los pasos anteriores, en el ya señalado orden de antes los primeros y después los posteriores: ajustando, retocando y modificando con el propósito de corregir aquello que no parezca adecuado. Aquí está uno de los puntos clave de un buen proceso de minería de datos: para su buena consideración hace falta, como se ha dicho, el concurso del experto en el problema y del diseñador; o, mejor aún, una decidida y franca colaboración entre ambos.

3.3.5. Aplicación

Y se llega a la aplicación real, de campo: de la que se obtiene una nueva evaluación, y con la que es posible realizar reinterpretaciones, a lo largo de toda la vida útil del sistema, obteniendo con ello un (mejor) conocimiento del problema que se está resolviendo; conocimiento del que, otra vez, pueden derivarse orientaciones para repetir el recorrido y la revisión del diseño; lo que, de nuevo, hace recomendable contar con la presencia del diseñador. Incluso aunque se obtenga ventaja del empleo de algoritmia de carácter adaptativo en diversos componentes del sistema completo de minería, lo que tiene particular valor cuando se trata un problema en un entorno de características temporalmente variantes, la ayuda del diseñador se requiere para adaptar lo adaptativo y posibilitar una eficaz actualización automática del sistema.

El conocimiento así obtenido es el resultado final de la minería de datos, lo que permite al experto la fundada toma de decisiones, que es, como anteriormente se ha dicho, aquello para lo que verdaderamente estamos bien dotados los humanos.
Como ilustración de este proceso, se ha recogido en el apéndice II un ejemplo de implantación paso a paso de una solución de minería de datos para resolver un problema/pregunta específico.

3.3. ALGUNAS REFLEXIONES

El carácter artificial de la minería de datos y sus difíciles bases analítico-computacionales despiertan no pequeña alarma y abundante rechazo, como ya se ha señalado en estas páginas, no sólo de quienes ven sus datos incluidos en las bases que se están tratando, sino hasta de sus potenciales usuarios. Insistir en cómo evitar los perniciosos efectos de estos miedos no es inapropiado, puesto que constituyen obstáculos iniciales no ya para la aplicación de la minería de datos, sino para el propio beneficio de clientes y expertos. En particular, no resultará vano recordar la necesidad del absoluto respeto a la legislación sobre protección de datos.
Es bien sabido que la oposición al cambio nace de la falta de incentivos y del miedo al riesgo. Ambas cosas actúan sobre clientes y sobre expertos: los primeros ven amenazada su intimidad y, para ello, hay que garantizar seguridad y confidencialidad; no aprecian así el beneficio que pueden obtener al ver personalizada su relación con la organización que considera sus datos. Los segundos temen a un instrumento, algo tan paradójico como el temor a las lentes correctoras de la visión cuando son necesarias, o al menos convenientes; además, cosa muy explicable, tampoco desean meterse en camisa de once varas.

Para salir de esta parálisis, la organización interesada es quien tiene que dar el primer paso: al fin y al cabo, es la que aspira al beneficio del proceso, ya que satisfacer y fidelizar a sus clientes y mejorar el rendimiento de sus empleados, a la vez contentándolos, son cosas que hacen falta para y desembocan en tal beneficio. Por eso la organización tiene que involucrarse en la implantación del proceso de minería de datos: para el empujón inicial (y también, como más adelante veremos en algún caso paradigmático, para evitar que los departamentos se vean sometidos a una confrontación de intereses). Y, para ese primer paso, la organización ha de conocer las ventajas esperables de la Minería. Para tal propósito no basta el esfuerzo de investigadores, diseñadores, consultores y proveedores: pese a la aparente circularidad de la proposición son los expertos el componente principal de esta sensibilización, que deben abordar antes de que el mercado haga saber que se va con retraso o, peor aún, que es demasiado tarde.

El experto tiene la llave, con riesgo casi inexistente, y teniendo en cuenta el consejo de investigadores, diseñadores, consultores y proveedores, para inducir a su organización a buscar ventaja en la minería de datos: si la organización lo acepta, la ganancia del experto es muy grande, en recursos disponibles y en aprecio de su actividad; y se crea un «círculo virtuoso» de intereses satisfechos de difícil ruptura. Con el apropiado uso de la minería de datos, que posibilita al experto extraer información previamente oculta, este experto adquiere un «sentido» adicional para observar la realidad: un «sentido» con limitaciones —como las tienen los propios sentidos humanos—, pero que suministra información valiosa, que, mediante el necesario aprendizaje, le permitirá adquirir y desarrollar nuevo conocimiento. Si no acepta actuar en esta dirección, y ya que es difícil admitir que la expansión de estas tecnologías se detenga por causa de resistencias ocasionales, el experto será el primer perdedor. Debe el

experto comprender que la única capacidad que le es estrictamente propia es la de tomar o proponer decisiones: tal capacidad, repetimos, puede ejercerse mejor y con menos riesgos si se incorpora este tan adecuado nuevo «sentido», colaborando con los que pueden enseñarle su buen empleo.

3.4. Y ALGUNOS ASPECTOS PRÁCTICOS

Tras repetir que es la organización interesada la que tiene que dar el primer paso, hemos de decir que no ha de hacerlo a ciegas: ha de definir claramente cuáles son los objetivos que se buscan en la aplicación de la minería de datos, y hacerlo en virtud de unos ciertos retornos esperados. Y el primer paso no ha de abarcar la totalidad del proceso: es frecuente, y hasta diríamos que recomendable, empezar por un proyecto piloto, con ayuda de un equipo técnico perteneciente a un proveedor, consultor u organización de I+D, pero proporcionando (al menos) dedicación de personas conocedoras del negocio. Además de que así se facilita la comprensión técnica del proceso y el diálogo entre diferentes tipos de componentes de un equipo de minería de datos, el usuario puede apreciar las características de la(s) herramientas(s) que se pone(n) en juego: su elección implica la de las técnicas de diseño y puede condicionar el despliegue de las tareas de minería en fases posteriores.

Un equipo completo de minería de datos incluye la presencia de expertos en el negocio, de consultores de negocio y de técnicas cuantitativas, de analistas y de diversas clases de programadores: son evidentes sus papeles, considerando la precedente descripción del proceso. No obstante, tales equipos resultan necesarios sólo para organizaciones de gran tamaño, que típicamente han de llevar a cabo simultáneamente varias aplicaciones de la minería de datos trabajando con bases de datos muy masivas. Ver-

siones reducidas de este equipo completo pueden ser más que suficientes en otros muchos casos: incluso la mínima (personas de negocios más técnicos) análoga a la antes citada para un proyecto piloto. Eso sí: en cualquier configuración ha de respetarse la idea fundamental de que una correcta minería de datos ha de suponer una buena conexión entre negocio y técnica.

Cerramos este punto con una reiteración: los datos reflejan el desarrollo del negocio tal como es y, por tanto, son la mejor fuente de información sobre el mismo; y no basta con observarlos, sino que hay que extraer esa información para posibilitar la adquisición de lo importante, el conocimiento sobre el negocio. Aun reconociendo que la adopción de la minería de datos requiere esfuerzos y cuidados, no debe olvidarse que, como quiera que han pasado los tiempos fáciles, tales esfuerzos y cuidados son imprescindibles para el avance del negocio, cuando no para su supervivencia.

4

APLICACIONES DE LA MINERÍA DE DATOS

4.1. UNA TIPOLOGÍA (PARCIAL) DE LAS APLICACIONES DE LA MINERÍA DE DATOS

Todo intento de elaborar una tipología exhaustiva de la minería de datos está condenado al fracaso: cualquier actividad observable requiere, para su comprensión, de estimaciones y decisiones, y esto es lo que permite la minería de datos; por tanto, si la actividad es registrable como datos y estos son abundantes (lo que ocurre en un enorme número de ocasiones), tenemos una posible aplicación de la minería de datos.

Optamos aquí por una tipología (harto) restringida, poniendo especial interés en aplicaciones relacionadas con diversos aspectos de los negocios, dado el potencial destino de este texto. Debemos excusarnos por presentar los tipos según una «doble entrada» —por sector y por aplicación—, pues nos hemos visto obligados a ello para evitar que un lector con intereses específicos haya de recorrer todo el listado so pena de perder información relevante.

4.1.1. Telecomunicaciones

- Detección de fraude en llamadas telefónicas.
- Gestión de la rotación («churn») de abonados.

- Segmentación de mercados en grupos de usuarios para márquetin y fidelización
- «Upgrading» de servicios.
- Análisis de flujos de datos de alta velocidad.
- Detección de fallos de red y seguridad informática.
- Segmentación (diversificación) de carteras de clientes.

4.1.2. Comercio y Márquetin

- Fidelización de clientes y relación con el cliente («Customer Relationship Management», CRM).
- Márquetin dirigido («targeting»).
- Estimación de la vida útil comercial de clientes.
- Prospección, análisis de mercado y análisis de cesta de la compra.
- Predicción de ventas.

4.1.3. Sector Farmacéutico y Sanitario

- Predicción de ventas de productos farmacéuticos.
- Ayuda al diagnóstico médico.
- Investigación farmacéutica: supervisión de cultivo industrial de antibióticos y descubrimiento de nuevos fármacos.
- Detección de fraude médico.
- Análisis de datos clínicos en hospitales.

4.1.4. Sector Administración Pública y Servicios

- Prevención del crimen y terrorismo.
- Investigación científica.
- Control y predicción de tráfico de vehículos.
- Predicción de flujos y optimización del turismo.
- Diseño de políticas de gobierno.

4.1.5. Sector Financiero

- Detección y control de fraude en el uso de tarjetas de crédito.
- Calificación de créditos hipotecarios y al consumo.
- Evaluación de «salud financiera» de entidades.
- Segmentación y fidelización de clientes.
- Detección de fraude en gestión y operaciones financieras.
- Segmentación (diversificación) de carteras de clientes.

4.1.6. Seguros

- Estimación de riesgos en la concesión de seguros.
- Análisis de rotación de clientes («churn»).
- Detección de fraude en seguros.

4.1.7. Industria y gestión empresarial

- Prevención de accidentes y gestión de alarmas (plantas petroquímicas, nucleares, etc.).
- Monitorización, control de procesos y diagnóstico automático de funcionamiento.
- Gestión y optimización de almacenes y organizaciones.
- Inteligencia de negocio.
- Diseño y manufactura.
- Gestión de conocimiento estratégica.
- Ayuda en la toma de decisiones.

4.1.8. Internet/Comercio electrónico/Textos

- Perfilado de usuarios para minería de uso web («web usage mining»).
- Ayuda a la navegación web y rediseño de sitios web para optimizar beneficios.

- Campañas de márquetin dirigido uno-a-uno («one-to-one») y anuncios inteligentes personalizados («banners»).
- Diseño de ofertas en comercio electrónico.
- Análisis automático de textos/filtrado de noticias.
- Descubrimiento de patrones de comportamiento en comercio electrónico.

4.2. EXAMEN DE ALGUNOS CASOS REALES

Pese a haber intentado, en páginas anteriores, formular advertencias y recomendaciones que ayuden a evitar problemas y errores y a obtener un buen rendimiento de los procesos de minería de datos, sabemos que no hay nada como la realidad como fuente para el aprendizaje: por ello exponemos algunos casos significativos, con la intención de que sus rasgos refuercen el buen entendimiento de los principios expuestos.

4.2.1. El programa «Customer First» de MCI

En 1992, MCI se encontraba en plena competencia con AT&T y Sprint en el mercado de las llamadas a larga distancia. El Departamento de Márquetin buscaba aumentar las altas ofreciendo directamente cheques cuando se producían éstas; mientras tanto, el Departamento de Ventas, tras el éxito del hoy bien conocido programa «Friends & Family» en 1991, quería mejorar la fidelidad de los buenos clientes identificándolos, analizándolos y ofreciéndoles incentivos «personalizados».
Desde el principio, Ventas se encontró con serias dificultades para alcanzar su objetivo: los datos sobre clientes estaban dispersos en diferentes silos y, además, mal gestionados: así, un abonado con dos números se consideraba como dos abonados independientes, un cambio de domicilio se interpretaba como una baja y un alta, etc.

Ventas se enfrenta a todas esas dificultades y lleva a cabo un primer proceso de minería de datos, obteniendo como resultados más relevantes:

- La rotación («churn») concluía mayoritariamente a los tres meses de producirse el alta, y a los dieciocho meses podía considerarse que el cliente estaba fidelizado; en función de lo cual, Ventas recomienda pasar del cheque por alta a incentivos «a plazo» (cheque a los tres meses, etc.).
- Los clientes fieles, en su mayoría, procedían del programa «Friends & Family» o/y estaban asociados a una empresa colaboradora (línea aérea, tarjeta de crédito, etc.).
- La identificación de los mejores clientes lleva a conocer que los que generaban ingresos superiores a 75 US $ al mes (500.000: el 5% del total) suponían, en su conjunto, el 40% de los ingresos, y que eran numerosos entre ellos los trabajadores en su domicilio, los que mantenían conferencias trasantlánticas, los viajeros frecuentes...; pero las opciones de personalizar quedaban muy limitadas porque subsistían importantes preguntas sin respuesta: los trabajadores en casa, ¿eran autónomos?; los conferenciantes trasantlánticos, ¿eran anglohablantes?; etc.

A la vista de estas incertidumbres, Ventas adopta una opción razonable: crear el programa «Customer First», en el que se asignaban a los buenos clientes gerentes de carteras personales, además de números 800, y se lanzaban campañas de planes de llamadas individualizadas a través de correo y telemárquetin segmentado, al tiempo que se aprestaba, con los datos que obtendría, a reevaluar las métricas y otros aspectos del previo proceso de minería... Pero no hay peor enemigo que el éxito parcial si no se tienen los pies fuertemente apoyados: el descenso de la rotación provocado por alguna de las medidas adoptadas produce daño al Departamento de Márquetin, que recibía

pluses en función del número de altas. Márquetin reclama ante los responsables corporativos hacerse cargo del programa: se le transfiere y, tras rebautizarlo como «Personal Thanks», lo mata.
Pese a que Ventas consigue mantener un programa «uno a uno» con el mejor 0,15% de los clientes, que generan el 7% de los ingresos, podemos, en consecuencia, concluir que la experiencia de minería de datos descrita constituye un fracaso. Pero obsérvese: no se debe a las dificultades intrínsecas (escasez y desorden de datos, etc.), sino que se produce porque la organización no se prepara para aprovechar los resultados. ¿Hace falta ahora insistir en el papel de los responsables de la organización?

4.2.2. La reducción de costes de campañas postales en Mellon Bank Corporation

Mellon Bank Corporation es una compañía estadounidense de servicios financieros, con sede en Pittsburgh. A mediados de los años noventa, y encontrándose ya en primera línea de la adopción de tecnologías avanzadas (gastaba unos 350 millones de dólares USA al año en computación), decidió acometer un proyecto de minería de datos encaminado a reducir costes de campañas de márquetin postal, centrándose para empezar en un caso concreto: ofertar a los clientes que tenían depósitos en la entidad y no lo eran de una línea de crédito con garantía en las propiedades inmobiliarias del titular (unos 210.000 casos) que acudiesen a dicha línea de crédito, recurriendo como datos etiquetados a los correspondientes a 40.000 casos que utilizaban ambos servicios y 5.000 más que habían rechazado el segundo (la etiqueta, aproximada, era la utilización o no de la línea de crédito). En realidad, se trataba implícitamente de aprender y evaluar las técnicas que pudieran reducir los grandes gastos que implicaban (e implican hoy, claro está) las campañas de márquetin pos-

tal no dirigido, cuyas tasas de respuesta se sitúan entre el 1% y el 2%, típicamente.
Un análisis estadístico preliminar llevó a considerar 120 variables como posiblemente relevantes: sociodemográficas (debe resaltarse que se excluyeron las conocidas como «discriminadoras inapropiadas»: edad, sexo, etc.), procedentes de la cuenta de depósitos y correspondientes al préstamo (éstas, para definir «subproblemas»). Tras ello, se utilizaron diseños de redes neuronales para una primera valoración de la probabilidad de respuesta de la población no etiquetada (en realidad, de su rango: de mayor a menor) ante diferentes tipos de oferta: en esta fase se realizó una selección de las variables relevantes. Seguidamente, y con el propósito de conseguir interpretar los resultados desde el punto de vista del negocio, se hizo una clasificación en casos positivos y negativos mediante un árbol de decisión.
Resulta muy ilustrativo revisar las conclusiones que extrajo la compañía de esta experiencia:

- Los resultados para el problema concreto que se atacó fueron más que satisfactorios: se conseguía llegar a más del 90% de los clientes que respondían con el 50% de los envíos, o a más del 50% con sólo el 20% de los envíos.
- La estructura del árbol de decisión resultante de la segunda fase permitió a los analistas preparar varias cartas de oferta distintas para diferentes subgrupos de clientes, al serles posible interpretar las características de éstos.
- También fue posible identificar factores críticos en este tipo de operaciones: así, tasas de interés superiores al 6,775% resultaban disuasorias.
- Mellon Bank pudo ahorrar importantes sumas prescindiendo de la compra a proveedores externos de registros de datos con variables que resultaban inútiles a efectos predictivos en estas aplicaciones.

- Finalmente, la compañía, a la vista de los esfuerzos que resultaron precisos para la preparación de los datos, enormemente mayores que los que se requirieron para la minería propiamente dicha, inició un proyecto estratégico para sistematizar la captura, el almacenamiento y los preprocesados básicos (no orientados a aplicación) de los datos de negocio, en el convencimiento de que así potenciaba grandemente sus posibilidades de incrementar beneficios.

4.2.3. El caso de Jubii: personalización en comercio electrónico

En un mes típico, Jubii, el portal de Internet más popular en Dinamarca, tiene 2,3 millones de visitas. De hecho, durante los últimos cinco años ha sido el principal portal de ese país, absorbiendo el 82% del mercado total. Es un subsidiario de Lycos Europa, que es parcialmente poseído por Lycos USA. Dicho portal ofrece una serie de servicios gratuitos: motor de búsqueda, correo electrónico, radiodifusión por internet, tienda, chat, etc.

Pero la popularidad no lo es todo: Jubii es un negocio, y como tal ha de mejorar sus cuentas de resultados mes a mes, en un terreno tan dinámico y competitivo como el de los portales web. Como los servicios ofrecidos son a coste cero para el visitante, su éxito/viabilidad comercial depende de los beneficios procedentes de anuncios («banners»), patrocinios, eventos Internet, boletín de avisos, entre otros. Debe, por tanto, atraer el máximo de anunciantes, y crear servicios de valor añadido por los cuales los usuarios estén dispuestos a pagar. Es necesario como paso previo comprender perfectamente qué es lo que los visitantes quieren comprar y/o visualizar.

Para aumentar su rentabilidad, el director de ventas Kasper Larsen lanzó un proyecto de perfilado de clientes, con el objetivo de crear perfiles de usuario, que ayudarían a

los anunciantes a un márquetin mejorado. Se han identificado y fijado los siguientes objetivos parciales:

- Aumentar el número y calidad de los clientes proporcionados a los anunciantes.
- Permitir a los anunciantes dirigirse a los visitantes web más proclives a la compra de sus productos.
- Extender su lista de visitantes, mientras retienen a los usuarios actuales.
- Crear servicios de valor añadido para atraer a suscriptores futuros.

Como primera fase se contactó con SPSS y ACsys, para que prestaran servicios de consultoría y productos de minería. Los datos de actividad de los usuarios fueron captados utilizando el componente «DoubleClick AdServer», y para el análisis de la información capturada, se empleó Clementine. Se midieron cuatro características para cada usuario registrado:

- Perfil de tipo de páginas visitadas en días laborables.
- Perfil de tipo de páginas visitadas en días festivos y fin de semana.
- Perfil de tiempo de uso (momento del día, duración), en días laborables.
- Perfil de tiempo de uso (momento del día, duración), en días festivos y fin de semana.

Cada uno de estos modelos se actualiza con cada acción del usuario, con lo cual está siempre disponible información actualizada de cada visitante al sitio Web. Antes de desplegar un nuevo anuncio («banner»), se consulta una base de datos de puntuación para encontrar el anuncio más apropiado para una determinada ubicación y usuario.
La expansión de este proyecto dentro del portal fue muy rápida, denotando buenas características de escalabilidad frente a grandes volúmenes de datos: al principio, se apli-

caban estas técnicas únicamente sobre 65.000 usuarios (clientes fidelizados en la sección «Jubii's Euro Investor», pero tres meses después del comienzo del proyecto ya se extendió a la totalidad de usuarios y secciones.

El éxito de un anuncio web («banner») se mide mediante la tasa de activación del mismo («clic-through rate», CTR), esto es, el número de personas que hacen «clic» sobre el mismo para acceder al sitio web del anunciante. Antes de la implantación del proyecto, la CTR en la sección «Euro Investor» era del 0,05%, tras la implantación, se produjo un incremento de entre el 30% y el 50%. Esto dio lugar a una mayor satisfacción de los anunciantes y, por tanto, una mayor fidelización de los mismos al portal Jubbi. Con objeto de hacer estos datos transparentes a los propios anunciantes, Jubii diseñó un nuevo interfaz gráfico para anunciantes, de modo que éstos pudieran comparar las prestaciones con el método antiguo frente al método incluyendo minería de datos, para demostrar que los efectos no eran casuales. Esta información es, a su vez, de gran interés para los anunciantes, pues al poder explorar efectos de respuesta en su mercado —incluso incluyendo la posibilidad de centrarse en grupos especiales, por secciones, en función del tiempo, etc.—, están más capacitados para un mejor diseño de sus campañas de márquetin, su adecuada planificación temporal o temporización, los objetivos que se pretende alcanzar, etc.

Como las agencias de medios realizan compras basándose en el índice CTR, esto implica directamente un aumento de ventas. El incremento observado en CTR se traduce, finalmente, en un aumento de las ventas entre el 10% y el 15%. Con respecto a los costes de implantación, cifras estimadas de negocio indican que el coste del proyecto se ha amortizado completamente en unos diez meses, lo cual hace totalmente viable y asumible el despliegue de estos servicios en casos reales.

Como posible trabajo de mejora futura, se ha observado que la información extra medida de las interacciones de

los usuarios proporciona a Jubii la posibilidad de establecer nuevos métodos más flexibles de facturación: por ejemplo, a los anunciantes se les podría cobrar por el número de veces que un visitante selecciona determinada página, por colocar anuncios en las franjas horarias más provechosas, etc., es decir, se abren nuevas formas de negocio y facturación a partir de los análisis llevados a cabo sobre los datos de usuarios.

El interés inicial sobre los mecanismos de anuncio mediante «banner» fue la forma práctica de demostrar la importancia de conocer los hábitos de los consumidores para optimizar los beneficios, pero es intención de Jubii extender el uso de estos mecanismos al resto de contenidos y servicios del portal, a fin de alcanzar una personalización total de lo ofrecido en el mismo, estableciéndose unas bases sólidas para el despliegue de mejores sistemas de gestión y creación de contenidos.

En el futuro, Jubii planea también monitorizar su tasa de rotación («churn») de usuarios mediante la generación de grafos utilizando datos de terceros, esto es, necesita construir una escena general del mercado, averiguando no sólo cifras absolutas de acceso a su portal, sino cifras relativas (incremento, decremento) de número de visitas a este tipo de portales.

4.2.4. ClearCommerce Corporation: reducción de riesgo y detección de fraude en operaciones en Internet

ClearCommerce Corporation es un líder mundial en soluciones para procesamiento de pagos y detección de fraude en operaciones realizadas a través de Internet, habitualmente haciendo uso de medios electrónicos de pago.

De acuerdo con un estudio del año 2002 del Grupo Gartner, uno de cada seis consumidores en Internet de los Estados Unidos fue víctima de un fraude en tarjetas de cré-

dito durante el año 2001. El 95% de las operaciones de compra en Internet se llevan a cabo mediante este medio de pago a crédito, y el fraude es 19 veces más frecuente en Internet que en establecimientos presenciales. Pese a los esfuerzos continuados por parte de comerciantes, intermediarios, emisores de tarjetas y cuerpos policiales para reducir dicho fraude, las cifras siguen siendo preocupantes y de hecho experimentaron un crecimiento entre los años 2000 y 2002.

ClearCommerce fue fundada en 1997 como un proveedor de servicios de pago a sitios web y, desde 1999, ha venido utilizando una aplicación denominada FraudShield™, que básicamente es un sistema basado en reglas que detecta potenciales operaciones fraudulentas y ayuda a los comerciantes a detectar transacciones sospechosas. Pero la mera detección no era suficiente para mantener en cifras razonables los casos de fraude; por lo cual se decidió complementar la aplicación con un módulo predictivo de calificación. El director de minería de datos en ClearCommerce, Daniele Micci-Barreca decidió incorporar un módulo de redes neuronales que ya conocía de aplicaciones análogas en otros ámbitos, para su uso como evaluación *a priori* del riesgo de las transacciones, concretamente la aplicación FraudAnalyzer de Clementine.

El sistema de decisión basado en tecnología de redes neuronales requería datos para el ajuste de sus parámetros libres, fase también denominada como de «entrenamiento», y para ello se recurrió a la gran cantidad de información almacenada por ClearCommerce a lo largo de sus muchos años de funcionamiento en el mundo de las ventas en Internet. Se procedió al análisis de millones de transacciones económicas, así como a decenas de miles de casos auténticos de fraude. Estos datos históricos recopilaban el historial de transacciones de miles de sitios web, de los que se habían recopilado tanto detalles de las operaciones (cantidad de la compra, momento del día, información de facturación, detalles de envío, etc.) como información de de-

voluciones o negaciones de transacción de operación de tarjeta de crédito. Una vez más, se puede intuir que las buenas prestaciones alcanzadas por el sistema se deben en buena medida a la importante y detallada cantidad de información recopilada con el paso de los años.

De este modo, el conocimiento previo sobre los determinados negocios, aportado por los directamente implicados y plasmado en el primitivo sistema de reglas, se complementó con un sistema de análisis de datos para la calificación de riesgo transaccional. Este último podía, en última instancia, permitir la extracción de reglas interpretables que se incorporaban al sistema de reglas predefinidas, lo que aumentaba la explicabilidad de las acciones o decisiones tomadas por el sistema.

Es difícil evaluar qué cantidad de operaciones fraudulentas ha podido evitar el sistema, pues las operaciones rechazadas nunca son analizables y, además, cada implementación se ajusta a los parámetros de cada cliente. Cada implantación es flexible, de modo que una vez en funcionamiento, cada usuario final puede modificar los parámetros oportunos para permitir un mayor porcentaje de transacciones aceptadas, a costa de un mayor riesgo de fraude o viceversa. No obstante, sí es posible comparar con variables medidas en términos globales, concretamente, en relación a las tasas medias de rechazo de operaciones en el sector, y el coste de revisión manual de órdenes de compra.

La tasa de rechazo de operaciones indica el porcentaje de peticiones que han sido devueltas al comerciante, debido a que el propietario de la tarjeta de crédito las rechaza, debido fundamentalmente a que eran fraudulentas (uso no autorizado de la tarjeta). La tasa estándar de rechazo en términos generales se sitúa entre el 0,7% y el 0,8%. En la mayor parte de los negocios que han implantado la solución de ClearCommerce, dicha tasa se ha reducido hasta el 0,1%.

En lo que respecta a la segunda variable, el porcentaje de órdenes revisadas manualmente, se ha observado que di-

cha cifra se veía reducida hasta el 40% aun conservando las cifras de rechazo antes mencionadas, lo que supone un enorme ahorro en costes de personal asociado a dicha tarea rutinaria de inspección manual.
Hemos visto cómo el uso de técnicas de minería de datos representa una nueva arma para luchar contra fenómenos como el fraude en transacciones en Internet que, de otro modo, experimentarían un crecimiento generalizado y sin control.

4.2.5. VISANET Brasil: detección de fraude en operaciones con tarjetas de crédito

VISANET, radicada en São Paulo (Brasil), es la mayor procesadora de tarjetas de crédito en su país, con casi tres millones de transacciones al día, en prestación del servicio a la casi totalidad de los bancos que operan en el territorio brasileño.
La prevención del fraude en el uso de tarjetas de crédito supone enfrentarse con un problema con las características típicas de la minería de datos: un volumen muy alto de datos que hay que manejar y necesidad de alta discriminación entre muestras que comparten un buen número de características. Además, la dificultad de la solución aumenta por el gran desequilibrio que existe entre las poblaciones correspondientes a operaciones correctas y operaciones fraudulentas. Desde el punto de vista operativo, los desafíos para implantar soluciones de éxito se ven incrementados aún más por el interés de obtener respuestas muy rápidas —preferiblemente en tiempo real— ante transacciones concretas, y por la ubicación del sistema de detección en un proceso obviamente crucial para el proveedor de servicios de medios de pago, las entidades bancarias y los usuarios, lo que obliga a satisfacer unos requerimientos de calidad de «software» y de seguridad y privacidad de información extremadamente exigentes.

Hace pocos años, VISANET estableció contacto con el Instituto de Ingeniería del Conocimiento (IIC), sito en la Universidad Autónoma de Madrid, que había empezado a trabajar en esta problemática en 1996, en un proyecto conjunto con IBM España. La evolución de estos trabajos dio posteriormente lugar a la creación del sistema Lynx, cuyas bases tecnológicas son de desarrollo propio.

El sistema Lynx combina la aplicación de módulos paramétricos, que permiten incorporar reglas que reproducen el comportamiento de analistas expertos, y modelos neuronales, que se entrenan con datos históricos propios de clientes, incluyendo un fichero de fraude comprobado; entrenamiento que se repite periódicamente para adaptarse a la dinámica del fraude y hacer frente a nuevas necesidades. Lynx asigna a cada operación con tarjeta un índice de riesgo entre 0 y 100, asimilable a la probabilidad de que la transacción sea fraudulenta. Ofrece así un bajo cociente entre número de casos que dan lugar a alerta y casos realmente fraudulentos, lo que permite a la entidad que lo implante un importante ahorro tanto por fraude evitado mediante el bloqueo temprano de las tarjetas cuanto en los propios costes del proceso de detección, al ser posible reducir el número de casos que requieren análisis experto.

VISANET Brasil adoptó Lynx (que también ha sido incorporado por otros bancos y proveedores de servicios de pago en España e Iberoamérica) hace tres años: el resultado ha sido el rechazo de transacciones fraudulentas que habrían ocasionado unas pérdidas directas de unos 20 millones de dólares; y, si se tienen en cuenta los saldos en riesgo de las cuentas asociadas, la pérdida evitada en fraude se cifra en unos 65 millones de dólares. En palabras de Antonio Machado Jr, Director Ejecutivo de Riesgo y Fraude de VISANET:

«No existe nada similar en el mundo, en que casi el 100% de las transacciones de un país son monitorizadas por un único sistema accesible a todos los emisores locales con

garantía de seguridad en la información de cada uno de ellos».

4.2.6. La segmentación de clientes de ENDESA

ENDESA es el mayor de los cinco grandes proveedores de electricidad en España, con el 40% de cuota de mercado, y el segundo distribuidor de gas; ofrece también otros productos y servicios asociados a la energía. Antes de la liberación del mercado de energía en 2003, la observación de que, en los dos años precedentes, el 25% de las grandes empresas (únicas en disponer de esta opción) cambiaban su proveedor llevó a ENDESA a concluir que era necesario conocer su mercado, personalizar sus ofertas y saber a qué clientes fidelizar, según palabras de Luis Miguel Muruzabal, Director de Márquetin.

ENDESA, que disponía de siete bases de datos masivas con diez millones de clientes, no había establecido variables relevantes para la subsegmentación de los tres segmentos tradicionales de clientes (hogares, empresas y grandes compañías). Los datos eran de calidad variable, no se analizaban por medios informáticos y se destinaban a la administración de clientes, con tiempo de respuesta lento.

La conveniencia de segmentar para objetivos de márquetin llevó a ENDESA a recurrir a un proveedor, que prestó su apoyo basándose en demostrar los beneficios de la minería de datos mediante un proyecto piloto. Se invirtió en herramientas de márquetin y se realizó la subsegmentación, con especial cuidado en el estricto cumplimento de la legislación sobre protección de datos, tanto en seguridad física como solicitando el consentimiento de los clientes para usar sus datos.

El proyecto identificó dos o tres subsegmentos dentro de cada uno de los segmentos tradicionales. Un destacado efecto ha sido que las divisiones de márquetin se organi-

zaron según estos planteamientos: lo que es prueba de que el resultado del proyecto se mostró beneficioso. Por ello dijo Luis Miguel Muruzabal:
«Tenemos diez millones de clientes, desde personas en áreas rurales alejadas hasta grandes corporaciones que utilizan un gran volumen de energía. Segmentar el mercado es una tarea difícil, pero el equipo de la compañía proveedora ha tenido las habilidades técnicas necesarias para ayudarnos a conseguir esta segmentación.»
Tras esta experiencia, ENDESA se concentra en la gestión de la relación con el cliente (CRM: «Customer Relationship Management») y dedica cada vez más recursos a este tipo de proyectos, por entender que son imprescindibles para la supervivencia de la compañía.

4.2.7. Diseñando un medicamento: el caso de deCODE genetics

Una de las más recientes aplicaciones de la biotecnología consiste en la búsqueda de formas variantes de genes que pueden estar directamente implicadas en el desarrollo de enfermedades humanas. Su detección y la posterior investigación de sus funciones posibilita el diseño de medicamentos específicamente dirigidos contra las funciones no deseadas, previniendo o tratando así las enfermedades derivadas. Esta aproximación terapéutica se empieza a denominar «medicina a la carta», posible hoy gracias, de un lado, a la secuenciación del genoma humano y, de otro, a la aplicación de la minería de datos para tratar los muchos miles de datos cifrados en las secuencias resultantes, realizando comparaciones entre personas aquejadas y personas libres de una cierta enfermedad (de la que haya sospecha de predisposición o concausa genética).
Un ejemplo de actividad en esta tan prometedora línea de trabajo es el proporcionado por la empresa deCODE genetics, que en febrero pasado publicó los resultados de un

estudio de los integrantes de cerca de 300 familias islandesas, entre los que había 700 personas que habían sufrido al menos un ataque cardíaco: el que esta población muestre elevados índices de consanguinidad permite detectar variaciones genéticas con mayor facilidad. La empresa también estudió los genes de dos grupos de 700 británicos, con y sin historial de ataques cardíacos.
Las comparaciones realizadas han permitido relacionar un gen, llamado ALOX5AP, con el riesgo de ataque cardíaco: deCODE genomics detectó una frecuencia anormalmente elevada de formas alteradas de este gen en personas con este problema (30% entre los islandeses y 15% entre el correspondiente grupo británico); formas alteradas que producen cantidades inusualmente elevadas de compuestos (leucotrienos) que favorecen la inflamación de la pared arterial, probablemente incrementando el riesgo de infarto cardíaco. Esta información ha permitido poner en marcha el desarrollo de medicamentos específicos para bloquear la actividad de estos leucotrienos en personas con esta alteración genética, y con ello intentar reducir su riesgo de sufrir un ataque cardíaco.

5

EL ESTADO ACTUAL DE LA MINERÍA DE DATOS

5.1. ASPECTOS OBSERVADOS

Son varios los factores que dificultan la cuantificación de la situación actual del sector en términos de volumen de negocio asociado directamente a actividades de minería de datos. Por una parte, los datos económicos de consultoras que realizan labores en ese ámbito o de empresas proveedoras de soluciones de minería de datos no suelen ser fácilmente accesibles, al considerarse información estratégica. Por otro lado, no es fácil obtener medidas cuantitativas de volumen de negocio asociado a procesos de minería de datos, ya que éstos suelen utilizarse como medidas complementarias para mejorar la productividad de los negocios, y no dan lugar a resultados económicos con implicación explícita en los procesos contables. Nos limitaremos, por tanto, en este capítulo, a la interpretación del estado del sector mediante observaciones indirectas.

Aunque resulta interesante analizar casos particulares de mejoras en procesos y beneficios obtenidos tras aplicar técnicas de minería de datos en determinados ámbitos, —tal y como se ha ilustrado en los casos de aplicación recogidos en la sección 4.2—, también es de interés disponer de una perspectiva global, objetivo que se pretende cubrir en este capítulo mediante la presentación de datos de preferencias de uso de herramientas de minería de da-

tos, estadísticas de uso por sectores, y análisis del comportamiento en el mercado de empresas que comercializan este tipo de herramientas y/o servicios.

5.2. PREFERENCIAS DE USO DE HERRAMIENTAS

Como quiera que no existe acuerdo entre las diversas fuentes sobre la presencia entre los usuarios de la minería de datos de las diversas herramientas, sería imprudente incluir en este documento datos detallados.
Así, si bien en la encuesta realizada por KDnuggets [5] (muestra reducida para asegurar su validez científica), Clementine, de SPSS, figura en el primer lugar con un 14% de presencia en 2003 (aunque la suma de herramientas SAS también llega al 14%, la de SPSS alcanza el 23%), la cifra que le asigna Giga Research es el 2%, y el porcentaje del mercado mundial (en ventas) en 2002 que atribuye IDC a las herramientas de SPSS no llega al 7%, pasando SAS del 36%. De modo que sólo cabe decir que ambas compañías tienen importante presencia, como también es reconocible la de IBM («Intelligent Miner»), Microsoft (SQL), Insightful Corp («Insightful Miner»), Computer Associates, Hitachi y otros. No es despreciable el número de casos en que se emplea código propio o desarrollos sobre herramientas generales, como Matlab (Mathworks).

5.3. ÁMBITOS DE APLICACIÓN

En lo que respecto a ámbitos de aplicación de técnicas de minería de datos y, también procedentes de una encuesta en KDnuggets, ilustramos en la siguiente tabla el porcentaje de usos por sector.

---

[5] http://www.kdnuggets.com/

	2001	2002	2003
Bioinformática/biotecnología/farmacéutico	15	14	16
Banca	12	13	13
Marketing directo	–	11	10
Detección fraude	10	11	9
Datos científicos	9	6	9
Seguros	6	6	8
Telecomunicaciones	10	8	8
e-comercio/web	14	10	5
Inversiones/gestión almacén	5	4	3
Otros	19	17	19

Tabla 1

Ámbitos preferentes de aplicación de herramientas de minería de datos, según encuesta realizada por KDnuggets. Los resultados se presentan ordenados atendiendo a los datos del año 2003.

El sector predominante durante el período bajo estudio ha sido el de bioinformática/biotecnología/farmacia, seguido muy de cerca por el de banca, manteniéndose aproximadamente en ambos casos los porcentajes a lo largo de los tres años. Se observa un declive importante en las aplicaciones en e-comercio y web, tal vez debido a las expectativas no satisfechas en ese ámbito de negocio en los últimos años, aunque el hecho de que actualmente se observa una tímida recuperación en el sector, especialmente en Estados Unidos, hace pensar en un resurgimiento a medio plazo de aplicaciones en el mismo.

5.4. PREFERENCIAS DE USO DE TÉCNICAS

Resulta también interesante analizar qué tipo de técnicas son las preferidas por los usuarios de minería de datos. En otra encuesta paralela, también procedente de KDnuggets,

se recoge la evolución de las preferencias de los usuarios (tabla 2)

	Votos 2002	% uso	Votos 2003	% uso	Cambio
Árboles de decisión	128	60,1%	120	56,6%	–5,8%
Agrupamiento	103	48,4%	93	43,9%	–9,3%
Estadística clásica	101	47,4%	92	43,4%	–8,5%
Redes neuronales	75	35,2%	71	33,5%	–4,9%
Regresión logística	75	35,2%	69	32,5%	–7,6%
Visualización	52	24,4%	55	25,9%	6,3%
Reglas	63	29,6%	42	19,8%	–33,0%
k-NN	42	19,7%	38	17,9%	–9,1%
Procesado textos	30	14,1%	30	14,2%	0,5%
Minería web	19	8,9%	29	13,7%	53,4%
Máquinas de vectores soporte (SVMs)		-	24	11,3%	—
Métodos bayesianos	24	11,3%	24	11,3%	0,5%
Análisis secuencial	27	12,7%	24	11,3%	–10,7%
Métodos híbridos	21	9,9%	23	10,8%	10,0%
Algoritmos genéticos	26	12,2%	12	5,7%	–53,6%
Otros	20	9,4%	22	10.,4%	10,5%

Tabla 2

Cambios en las preferencias de técnicas de minería de datos. Los resultados se presentan ordenados atendiendo a los datos del año 2003.

Atendiendo a estos datos, la técnica de mayor uso y difusión es la de árboles de decisión, siendo utilizada por el 60% de los encuestados, posiblemente debido a su relativamente bajo coste computacional y buena escalabilidad, así como la posibilidad de interpretación final por parte

de usuario. No obstante, como se discute en el apéndice I, esta técnica no está exenta de inconvenientes. Le siguen de cerca las técnicas de agrupamiento, de estadística clásica y de redes neuronales. Como otros datos destacables, observamos el espectacular descenso del uso de los algoritmos genéticos, probablemente debido a las limitaciones de escalabilidad frente a grandes volúmenes de datos y problemas complejos. Tampoco es despreciable el descenso de la técnica de reglas, el 33%. Como caso de gran variación positiva, observamos el espectacular crecimiento de las técnicas de minería web, seguramente empujadas por el crecimiento exponencial de la información disponible, y el mayor y más asequible acceso de ésta al público en general, lo que motiva la optimización de los negocios web en general, pero que aun así sólo copan el 14% del total. Finalmente, es necesario resaltar el importante crecimiento de uso de una técnica emergente, las máquinas de vectores soporte, que en muy poco tiempo ya se equiparan a técnicas históricamente asentadas como las bayesianas.

5.5. UN DETALLE SOBRE LA EVOLUCIÓN DE LA MINERÍA DE DATOS

Observar en detalle la evolución de uno de los proveedores de referencia de herramientas de minería de datos como es SPSS, nos puede resultar altamente informativo acerca de la salud del sector. Podemos quedarnos con datos puntuales correspondientes a resultados económicos de SPSS en el tercer cuarto del año 2003, y se puede hablar de unos beneficios de 52 millones de dólares, concretamente las ventas de productos de minería se incrementaron el 26% con respecto al mismo período del año anterior, aumento cuantificable en unos dos millones de dólares. Estas cifras son indicativas de que el sector de la minería se encuentra en buenas condiciones y con buenas

perspectivas de crecimiento en los próximos años. Además de una fotografía instantánea, interesa sobremanera la observación de la evolución temporal de este tipo de negocios. En la siguiente figura, se ha representado la evolución en bolsa (Nasdaq) durante los últimos diez años de importantes empresas del sector TIC [6] como SPSS, Nokia, Cisco e IBM.

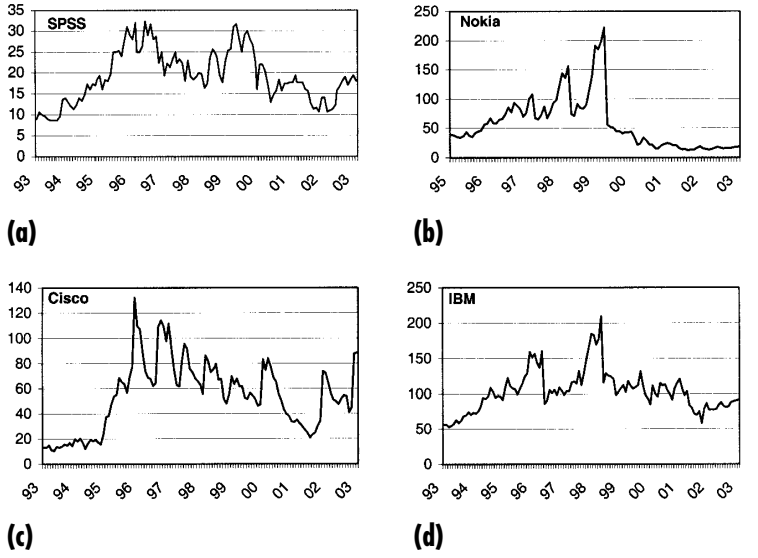

Figura 6

Evolución en bolsa (Nasdaq) durante los últimos diez años de importantes empresas del sector TIC: (a) SPSS, (b) Nokia, (c) Cisco y (d) IBM.

Se observa cómo SPSS e IBM, ambas compañías con activos importantes en minería de datos, han resistido relativamente bien el declive de los años 2001-2002,

[6] Tecnologías de la Información y las Comunicaciones.

de claro impacto negativo especialmente en el sector TIC. Hasta cierto punto, es posible inferir que la minería de datos es un ámbito de negocio relativamente robusto frente a «desastres» económicos, tal vez debido a que en tiempos de crisis, la mayor parte de las empresas acuden con premura a este tipo de soluciones que les permitan ganar la suficiente competitividad como para permitir su supervivencia. Para resumir este comportamiento, podríamos añadir a las frases recogidas en el punto 3.1 la de *«las empresas se acuerdan de Santa Bárbara cuando truena».* Adicionalmente, en épocas de bonanza económica —como la disfrutada hace años por las empresas tecnológicas («boom» de las «.com») o la que aparentemente se empieza a vivir en 2004—, tampoco faltan las fuentes de negocio, pues también son muchas las empresas que aprovechan los excedentes o buenas expectativas de resultados para mejorar sus procesos y optimizar sus negocios mediante la aplicación de minería de datos.

En lo que respecta a previsiones para los próximos años, debemos tener en cuenta hechos como que en los próximos tres años, la humanidad generará muchos más datos que los acumulados a lo largo de toda su historia, que cada veinte meses se duplicará la cantidad de información en el mundo, que la previsión de tráfico diariamente transmitido en 2007 por Internet será equivalente a 64.000 bibliotecas del Congreso de los Estados Unidos (estudio de IDC).[7] De modo complementario, atendiendo a aspectos puramente económicos, las previsiones para el futuro cercano sobre el sector TIC son halagüeñas; Gartner Group [8] pronostica explícitamente una sólida recuperación de una serie de sectores tecnológicos, entre los que se encuentran los de gestión de contenidos, minería de datos, inteligencia de negocio y gestión del conocimiento. Si a este creci-

[7] IDC Analyze the future. http://www.idc.com/
[8] http://www3.gartner.com/

miento en el volumen de datos acompañamos cifras económicas asociadas al análisis y procesado de los mismos, el volumen resultante de negocio alrededor de la minería de datos puede llegar a ser tremendamente importante, sobre todo cuando la adecuada extracción de información puede llegar a ser vital para un negocio desde el punto de vista de competitividad.

6

SOBRE OPORTUNIDADES Y OBSTÁCULOS

6.1. RECORDANDO LOS OBSTÁCULOS PRIMARIOS

De los obstáculos primarios que se oponen a la implantación de la minería de datos hemos hablado ya: las indecisiones y miedos de expertos y de clientes del negocio, así como la incomprensión y la pasividad de las propias organizaciones y la ciega irrupción de los técnicos en mundos desconocidos. Obstáculos también son para el buen aprovechamiento de las íntegras posibilidades de la minería de datos los señalados mediante frases de uso común como dificultades generales en el propio desarrollo del proceso. Pero aquí no pretendemos repetir, ni resumir siquiera, lo allí dicho: deseamos referirnos a otros obstáculos, digamos secundarios, que son los que, vencidos los primarios, siguen obscureciendo el camino hacia beneficios todavía mayores de la minería de datos: las oportunidades que se derivan de un planteamiento creativo de los correspondientes procesos.

6.2. LA MINERÍA DE DATOS «CREATIVA»

Lo nuevo y a la vez valioso, es decir, el resultado de la actuación creativa, se considera crucial para el progreso de los negocios: desafortunadamente, pese a ello no se pro-

picia tal modo de actuación... Una discusión sobre la creatividad, los obstáculos para que se manifieste y cómo potenciarla no cabe en esta páginas. Pero al menos un apunte que permita entrever las oportunidades de aumentar el provecho de la minería de datos mediante planteamientos creativos no debe omitirse.

Se ha dicho que la minería de datos busca dar respuestas a preguntas importantes para el ámbito o negocio en que se aplica: de hecho, un método directo es la formulación de preguntas, y su única limitación está precisamente en la capacidad de elegirlas. Y ya decía Pablo Picasso que los ordenadores son inútiles porque sólo saben dar respuestas... Si se acierta con las preguntas, sin imponer innecesarias restricciones, adentrándose en aspectos no completamente dirigidos a un objetivo específico, actuando con tranquila libertad, construyendo las interrogaciones de modo verdaderamente explorativo, las posibilidades se multiplican. No podremos exponer aquí nada más que unos cuantos ejemplos: creemos que serán suficientes para ilustrar las posibilidades adicionales de la minería de datos.

La cuestión de la relevancia de los datos se aprecia como fundamental por expertos y técnicos. Pero pocas veces surge la pregunta: *¿Qué otros datos podrían ser relevantes para mi problema?* Y así, se renuncia a examinar algunos ya disponibles o a incorporar otros fácilmente accesibles, y se cae en un estancamiento rutinario.

Ejemplo de respuesta que involucra datos disponibles: aun sabiendo que en el movimiento está la verdadera información —con él se desarrollan los sistemas nerviosos en los seres vivos—, en muchas ocasiones se incluye un «saldo medio» como única variable en un proceso de Minería. El saldo varía: no considerar que en esa variación hay información valiosa supone limitar indebidamente la potencia del proceso. Supongamos que las oscilaciones del saldo son infrecuentes y pequeñas respecto al valor medio: ¿no se deduce que el cliente probablemente dispone de otros recursos (en particular, si hay una nómina domiciliada)?

Supongamos que las variaciones del saldo tienen una distribución temporal absolutamente regular: ¿no implica que el cliente ha planificado cuidadosamente la gestión de esos fondos? Supongamos que sólo esporádicamente se producen retiradas de cantidades apreciables, tras un largo perfil creciente del saldo; ¿no se trata de un indicador de oportunidad?... Naturalmente, consideraciones análogas puedan hacerse sobre la dinámica de las llamadas telefónicas, de las compras en una gran superficie, etc., etc. Ejemplo de datos que se pueden conseguir con facilidad: ante una solicitud de crédito, ¿sería relevante saber qué otras cosas, aparte de las contenidas en un formulario, desea hacer notar el solicitante? Está claro que, ofreciendo esa posibilidad, el cliente no sólo se sentirá más apreciado, sino que puede suministrar datos de carácter personal —incluso emocionales— que no proveería de otro modo. Y no creemos que nadie discuta hoy la importancia de los aspectos emocionales en la conducta de las personas.

Abandonemos ya los ejemplos relativos a la primera de las preguntas que hemos propuesto, pese a que podría extenderse la lista hasta ocupar muchísimas páginas. Consideremos otra pregunta: *¿Es posible aprender algo más que lo relativo al objetivo específico de un proceso de minería de datos?*

Como inicio de los ejemplos correspondientes, volvamos al último anterior. Si un cierto número de clientes a los que se les pide que añadan lo que consideren apropiado destaca un cierto factor del que disponemos (aunque sea ocasionalmente) y no hacemos uso, y si comprobamos su relevancia: ¿no cabe la posibilidad de sistematizar su obtención y consideración? Si en un proceso que incluye la entrevista o la cumplimentación de un cuestionario se observa *a posteriori* que ciertos campos son irrelevantes y otros no, ¿no se puede revisar el cuestionario con ventaja? Ítem más: ¿no sería posible entrenar a los entrevistadores para que las entrevistas proporcionasen mejores datos y, consiguientemente, mayor información?

Los anteriores son ejemplos inmediatos del «metaaprendizaje» que se puede derivar de procesos de minería de datos: no nos parece preciso insistir en que también es posible encontrar muchas otras respuestas acerca de lo que la organización y sus componentes pueden conocer sobre lo que hacen y lo que deberían, o no deberían, hacer.

No se acaban aquí las preguntas y los ejemplos de respuestas: ante la cuestión *¿Dónde hay oportunidad de aplicar con provecho la minería de datos?* surgiría una lista interminable; pero, si se considera como requisito la novedad, cabe pensar en la prevención en los sistemas de Salud, en la proactividad en la atención al ciudadano, en la planificación de la actuación de las fuerzas de seguridad, en... Aquí el lector añadirá casos que nosotros no imaginamos siquiera.

También hay preguntas de carácter (aparentemente) muy técnico que inducen respuestas de alto valor: así la interrogación *¿Cómo puedo mejorar la fiabilidad de un proceso de segmentación?* podría llevar a contestarse que comprobando la consistencia de los segmentos: por ejemplo, intentando predecir características de un segmento a partir de datos de otros. Lo anterior no sólo posibilita apreciar inconsistencias: también influencias y mecanismos de acción-reacción cuyo conocimiento puede ser decisivo.

Invitar a un comportamiento creativo —una vez superadas las primeras barreras que se alzan ante la minería de datos— no es una mera cortesía: es una receta para el verdadero éxito. Los humanos empleamos para nuestros procesos mentales tanto mecanismos deliberativos cuanto mecanismos tácitos o intuitivos, y ahí está nuestra ventaja.[9] Combinarlos en la exploración que suponen los procesos de minería de datos posibilita ganar profundidad, comprensión... y conocimiento.

---

[9] Queremos aprovechar el momento para explicitar nuestra firme convicción de que también estará la de los expertos que admitan manejar técnicas «inexplicables» y reflexionar sobre sus resultados... No hay que despreciar radicalmente lo que, con cierto abuso, podría llamarse la «intuición» de las máquinas.

7

RELACIÓN DE PRESTADORES DE SERVICIOS

7.1. CENTROS DE I+D+i

7.1.1. Instituto de Ingeniería del Conocimiento (IIC)

El Instituto de Ingeniería del Conocimiento (IIC) es un Centro Tecnológico de Innovación que focaliza su actividad en el campo de la gestión del conocimiento, principalmente en dos áreas: «Capital Humano», relacionada con el desarrollo de sistemas para el ámbito de los recursos humanos, y «Minería de Datos y Conocimiento», centrada en la detección de patrones de comportamiento para el desarrollo de sistemas de detección de fraude.

Dirección postal:	UAM - Cantoblanco, Escuela Politécnica Superior Edificio B, 5.ª planta 28049 Madrid, España
Correo electrónico:	iic@iic.uam.es
Teléfono:	+34 91 497 23 23
Fax:	+34 91 497 23 34
Página web:	http://www.iic.uam.es/

7.1.2. Grupo de Tratamiento de Datos en la Universidad Carlos III de Madrid (GTD-UCIIIM)

La I+D del GTD-UCIIIM se centra en el desarrollo de nuevos algoritmos neuronales y evolutivos para su aplicación en problemas de minería de datos orientada fundamentalmente a aspectos de negocio. El GTD mantiene una actividad de I+D en temas de concepción y diseño de algoritmos neuronales y evolutivos, como son la propuesta de nuevos objetivos, el desarrollo de versiones «on-line», la combinación de aprendizaje y evolución y la interpretación de la actuación de los algoritmos. Simultáneamente, lleva a cabo la aplicación de los resultados a problemas reales de márquetin y negocio, financieros, de gestión de la información y de diagnóstico clínico e industrial, y, secundariamente, de comunicaciones y de tratamiento de señales. Dispone de capacidades de diseño y desarrollo de nuevos algoritmos para minería de datos, además de capacidad para formación y consultoría sobre el tema.

Dirección postal:	Departamento de Teoría de la Señal y Comunicaciones Universidad Carlos III de Madrid Avda. Universidad, 30 28911-Leganés, Madrid, España
Correo electrónico:	arfv@tsc.uc3m.es, navia@tsc.uc3m.es
Teléfono:	+34 91 624 99 23
Fax:	+34 91 624 87 49
Página web:	http:/www.uc3m.es

7.1.3. Grupo MIP: programación inductiva multiparadigma

Grupo de investigación centrado en el entrenamiento de modelos comprensibles bajo distintos paradigmas: programas de lógica funcional, árboles de decisión y listas de decisión. Sus áreas de investigación son: programación in-

ductiva multiparadigma, aprendizaje máquina y minería de datos, descubrimiento de conocimiento, análisis ROC, aprendizaje sensible al coste y evaluación de modelos para aplicaciones de ayuda a la decisión, así como el aprendizaje de modelos declarativos a partir de los datos.

Dirección postal:	Grupo MIP Departamento de Sistemas Informáticos y Computación Universidad Politécnica de Valencia Camino de Vera s/n 46022 Valencia, España
Correo electrónico:	jorallo@dsic.upv.es
Teléfono:	+34 96 387 73 50
Fax:	+34 96 387 73 59
Página web:	http:/http://www.dsic.upv.es/~flip

7.2. CONSULTORES Y DESARROLLADORES

7.2.1. Daedalus

Daedalus es una compañía fundada en 1998 como «spin-off» de la UPM y la UAM. Actualmente tiene 25 empleados y tiene un capital 100% privado. Sus actividades se centran en el sector de Tecnologías de la Información Avanzadas, siendo las áreas principales de actividad: tecnología lingüística para recuperación de información, minería web, optimización de sistemas complejos, minería de datos y tecnologías inalámbricas, gestión del conocimiento, extracción y recuperación de la información, categorización de textos y perfilado de usuarios.

Dirección postal: *Central*
DAEDALUS, S. A.
C/ López de Hoyos 15, 3.°
28006 Madrid, SPAIN

Departamento técnico
DAEDALUS, S. A.
Centro de Empresas «La Arboleda»
Carretera N-III, km 7,300
28031 Madrid, SPAIN
Correo electrónico: info@daedalus.es
Teléfono: +34 91 332 43 01
Fax: +34 91 331 97 40
Página web: http://www.daedalus.es/

7.2.2. ISOCO, S.A.

ISOCO (Intelligent Software Components, S.A.) es una empresa fundada en 1999 por un grupo de investigadores procedentes del Instituto de Investigación en Inteligencia Artificial, perteneciente al Consejo Superior de Investigaciones Científicas (CSIC). ISOCO crea y desarrolla constantemente soluciones innovadoras basadas en las más avanzadas tecnologías y en técnicas de Inteligencia Artificial, comprendiendo soluciones para aprovisionamiento estratégico, agregación inteligente, gestión del conocimiento, gestión de documentos oficiales y certificaciones, personalización en tiempo real, o visualizaciones de Internet.

Dirección postal: *Barcelona*
Alcalde Barnils, 64-68
Edificio Testa - bl. A
08190 Sant Cugat del Vallès
Teléfono: 93 567 72 00
Fax: 93 567 73 00

Madrid
Francisca Delgado, 11 - 2.°
28108 Alcobendas, Madrid
Teléfono: 91 334 97 97
Fax: 91 334 97 99

	Valencia Edificio Trade Center Profesor Beltrán Báguena, 4 of. 107 46009 Valencia
Teléfono:	96 346 71 43
Fax:	96 348 28 94
Correo electrónico:	isoco@isoco.com
Página web:	http://www.isoco.es

7.2.3. Meta4 Spain, S.A.

Meta4 es un proveedor importante de soluciones para la gestión y el desarrollo del capital humano e intelectual a nivel mundial. Fundada en 1991, Meta4 cuenta con más de 900 clientes en 18 países y sus soluciones gestionan 4,5 millones de personas en tres continentes. Meta4 cuenta con más del 45% de cuota de mercado en España (Estudio IDC, 2003). Las oficinas de la compañía, con 424 empleados, están ubicadas en Barcelona, Buenos Aires, Lisboa, Madrid, México DF, Santiago de Chile y París. Recientemente ha sido adquirida por el grupo Adonix.

Dirección postal:	Centro Europa Empresarial Edificio Roma - C/ Rozabella, 8 28230 Las Rozas, Madrid, España
Correo electrónico:	infmarketing@meta4.com
Teléfono:	+34 91 634 85 00
Fax:	+34 91 634 84 80
Página web:	www.meta4.com

7.2.4. Cognodata Consulting

Cognodata Consulting ofrece servicios de consultoría sobre optimización de campañas comerciales, captación se-

lectiva de clientes de alto valor, aumento de la rentabilidad mediante la maximización de ventas cruzadas, segmentación y priorización de nuevos mercados, valoración del nivel de calidad de datos, depuración y enriquecimiento de datos, análisis de oportunidades de creación de valor mediante análisis estratégico de datos, formación en análisis de datos con tecnología de minería de datos, apoyo en la creación de departamentos internos de análisis de datos y colaboración puntual en la construcción de modelos de minería de datos.

Dirección postal:	C/ Lagasca, 120, oficina 4 28006 Madrid, España
Correo electrónico:	info@cognodata.com
Teléfono:	+34 91 411 63 15
Fax:	+34 91 563 63 83
Página web:	http://www.cognodata.com/

7.2.5. IONE Consulting

Empresa de servicios de consultoría sobre minería web, especialmente para tareas de construcción de perfiles de audiencia demográfico y psicográfico, análisis del tráfico web, estudios del comportamiento de los usuarios (minería de uso de la web), para tareas de mejora del servicio, márquetin y rendimiento del sitio web, todo ello mediante la aplicación de técnicas de minería sobre los ficheros de registro de acceso almacenados en los servidores.

Dirección postal:	C/ Juan de la Cosa, 2, local 2H 29600 Marbella, Málaga, España
Correo electrónico:	info@ione.es
Teléfono:	+34 952 867 359
Fax:	no disponible
Página web:	http://www.ioneconsulting.net

7.2.6. Sigma Consultores Estadísticos

Sigma Consultores Estadísticos es una empresa especializada en el análisis cuantitativo de datos y proporciona servicios de consultoría, de diseño e implementación de todo tipo de investigaciones cuantitativas y ofrece la posibilidad de desarrollar herramientas de software a medida.

Dirección postal:	C/ Miguel Servet, 88, 5-1 50013 Zaragoza, España
Correo electrónico:	sigma@consultoresestadisticos.com
Teléfono:	+34 976 42 82 94
Fax:	no disponible
Página web:	http://www.consultoresestadisticos.com/

7.2.7. CHS Data Systems

Empresa para el asesoramiento, desarrollo e implementación de soluciones basadas en minería de datos para gestión de la relación del cliente («Customer Relationship Management», CRM), ofreciendo también consultoría en servicios informáticos y soluciones CRM integradas, cubriendo todo el campo relacionado con la gestión de los clientes —desde las soluciones operativas (Inteligencia de Negocio y minería de datos— hasta la completa integración de todos los procesos y sistemas.

Dirección postal:	*Central CHS S.L.* Avda. Ricardo Soriano, 12 Edificio Marqués de Salamanca 1.ª planta, oficinas 2 y 3 29600 Marbella, Málaga, España
Teléfono:	+34 952 76 66 80
Fax:	+34 952 86 81 64
	Madrid C/ José Abascal, 44, planta 4.ª 28003 Madrid, España

Teléfono:	+34 91 442 48 89
Fax:	+34 91 442 48 89
	Valencia C/ Jordi de Sant. Jordi, 10, 4.ª 46950 Xirivella, Valencia, España
Teléfono:	+34 963 832 993
Fax:	+34 952 868 164
Correo electrónico:	info@chs.es
Página web:	http://www.chs.es

7.2.8. Atos Origin

Atos Origin es un suministrador importante en consultoría en Tecnologías de la Información, integración de sistemas y servicios de redes e infraestructuras tanto para la industria energética como para el sector público y los mercados de telecomunicaciones y finanzas

Dirección postal:	C/ Albarracín, 25 28037 Madrid, España (Otras 5 sedes en Valladolid, Vigo, Barcelona, Sevilla y Bilbao; direcciones en http://www.schlumberger-sema.es/locations.htm)
Correo electrónico:	buzon-cliente@madrid.sema.slb.com
Teléfono:	+34 91 440 88 00
Fax:	+34 91 754 32 52
Página web:	http://www.sema.atosorigin.com/ Antigua URL en España, todavía activa: http://www.schlumbergersema.es/

7.2.9. Deloitte Consulting

Empresa de consultoría que proporciona soluciones integrales de gestión de empresa, capital humano, estrategia

y operación, e integración tecnológica. Desarrolla servicios de minería tales como CRM, gestión de contenidos, y márquetin inteligente.

Dirección postal:	*Madrid* Torre Picasso, planta 34 Plaza Pablo Ruiz Picasso, s/n 28020 Madrid, España
Teléfono:	+34 91 335 08 00
Fax:	+34 91 555 67 84
	Barcelona Moll de Barcelona, s/n World Trade Center Edificio Sur - 7.ª planta 08039 Barcelona, España
Teléfono:	+34 93 508 84 84
Fax:	+34 93 508 84 85
Correo electrónico:	No disponible, pero sí una página de contacto electrónico: http://www.dc.com/ContactUs/
Página web:	http://www.dc.com/

7.2.10. Soluziona

Empresa de soluciones integrales de consultoría en el ámbito de la gestión empresarial y las Tecnologías de la Información.

Dirección postal:	Paseo de la Habana, 101 28036 Madrid
Teléfono:	+34 91 210 20 00
Fax:	+34 91 344 03 86
	Otras muchas oficinas en Madrid y otras ciudades españolas, listado completo disponible en: http://www.soluziona.es/htdocs/areas/soluziona/empresa/oficinas/espana.shtml

Correo electrónico: No disponible, pero sí una página de contacto electrónico: http://www.soluziona.es/htdocs/global/utilidades/contacta/index.shtml
Página web: http://www.soluziona.es/

7.2.11. Indra

Indra es una compañía española con alta presencia en Tecnologías de la Información y Sistemas de Defensa. Mediante el Centro de Soluciones Business Intelligence asesora a los clientes sobre el almacenamiento, transformación y utilización de la información dentro de la organización. Sus soluciones y servicios están basados en Modelos de Gestión (Scorecard, Dashboard, ABC/ABM...) y Análisis (Query & Reporting, OLAP, Data Mining...) soportados por plataformas y productos de software.

Dirección postal: Avda. de Bruselas, 35
28108 Alcobendas (Madrid)
Teléfono: +34 91 480 50 00
Fax: +34 91 480 50 57
Correo electrónico: bi@indra.es
Página web: http://www.indra.es

7.3. PROVEEDORES

7.3.1. IBM

Empresa de primera fila en todo el mundo en Tecnologías de la Información y las Comunicaciones, que ofrece servicios de consultoría, y comercializa alguna de las herramientas de minería de datos más conocidas, como «DB2 OLAP Server», «Intelligent Miner for Data», o «Intelligent Miner for Text».

Dirección postal:	C/ Santa Hortensia, 26-28 28002 Madrid
Teléfono:	+34 91 397 66 11
Fax:	+34 91 519 39 87
Correo electrónico:	No disponible
Página web:	http://www.ibm.com/es/

7.3.2. SPSS Ibérica

Empresa con más de treinta años de experiencia en herramientas analíticas para minería de datos y análisis estadístico, con productos como Clementine, AnswerTree, LexiQuest o CFO Suite.

Dirección postal:	Plaza de Colón, 2 Torres de Colón, torre II, planta 16 28046 Madrid
Teléfono:	+34 902 123 606
Fax:	+34 91 308 35 21
Correo electrónico:	sales@spss.com
Página web:	http://www.spss.com/es/

7.3.3. SAS

Empresa fundada en los Estados Unidos hace veinticinco años y presente en muchos países. Es una compañía experta en soluciones de minería de datos, con herramientas como «Enterprise Miner» o «SAS Text Miner».

Dirección postal:	SAS Institute, S.A. Avda. Manoteras, 44 28050 Madrid
Teléfono:	+34 91 200 73 00
Fax:	+ 34 91 200 73 01
Correo electrónico:	sas@spn.sas.com
Página web:	http://www.sas.com/spain/

7.3.4. The Mathworks

Empresa experta en software científico y de simulación a nivel mundial. Aunque no específica de minería de datos, su principal herramienta, Matlab/Simulink, es muy usada para llevar a cabo tal tipo de tareas.

Dirección postal:	C/ París, 179-181, 1°, 2ª A 08036 Barcelona
Teléfono:	+34 93 362 13 00
Fax:	+34 93 200 95 56
Correo electrónico:	info@mathworks.es
Página web:	http://www.mathworks.es

APÉNDICE I LAS TECNOLOGÍAS PARA (EL TRATAMIENTO EN) LA MINERÍA DE DATOS

De acuerdo con lo indicado en el texto principal, se dedica este apéndice a una somera presentación de técnicas, más una exposición de sus ventajas e inconvenientes, sin aspiración alguna de exhaustividad: en particular, se omiten algunos procedimientos aún en desarrollo que, por ello, no han alcanzado presencia significativa en la minería de datos (como son las redes bayesianas, los filtros de partículas, las máquinas de núcleos y vectores soporte, etc.), así como otros cuyo principal empleo se encuentra en el diseño y la optimización, aunque eventualmente puedan utilizarse para decisión o estimación (algoritmos genéticos, evolutivos, sociales, etc.). Se completará el apéndice con una breve revisión de las tendencias más importantes, así como con unas palabras sobre las tres clases de herramientas (comerciales específicas, comerciales de propósito general y «software» de elaboración «ad hoc») a las que se puede recurrir para implementar el uso de las anteriormente citadas tecnologías.

TÉCNICAS

Visualización

Tradicionalmente, se entiende como visualización la representación de algunas de las variables que constituyen los

datos en prácticamente cualquier forma fácilmente comprensible: como nubes de puntos, histogramas, mapas, grafos, etcétera. Hasta pueden incluirse aquí métodos que se tienen como propios de otras técnicas que no son consideradas estrictamente como minería de datos: así, las representaciones jerárquicas (en ramificaciones consecutivas según el comportamiento de sucesivas variables) o los cubos, que son procedimientos típicos de lo que se conoce como procesamiento analítico «On Line» («On Line Analytical Processing», OLAP). Un cubo representa en cada una de sus tres dimensiones una variable (habitualmente temporal, espacial y de negocio), y en los subcubos resultantes el número de casos registrado: véase la Figura AI.1, que resulta autoexplicativa. Además de la obvia limitación del número de dimensiones visualizable, el empleo de las técnicas de visualización en este modo tradicional, sobre los datos, tiene de cierto una potencia muy limitada: sólo resaltará relaciones llamativas entre las variables que se están observando, lo que si bien puede tener importancia cuando la tienen dichas relaciones, y en todo caso puede ayudar al observador a comprender el problema que se considere, no permite apreciar relaciones de gran complejidad. No obstante, no hay que olvidar que el sistema visual humano está muy bien preparado para percibir información.

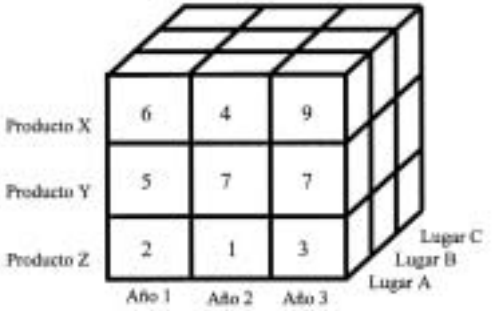

Figura AI.1:

Cubo de datos representando volumen de ventas: su examen se ha de hacer recorriendo las capas en la dimensión perpendicular al papel.

Limitaciones:

- potencia muy limitada
- puede producir equívocos

Ventajas:

- gran comodidad
- fácil interpretación

Recomendaciones de uso:

- como técnica exploratoria y auxiliar para y con otras

Preguntas

Como en el caso de la visualización, las preguntas no se suelen considerar estrictamente incluidas entre las técnicas de minería de datos, sino como una herramienta para examen exploratorio de los datos disponibles: mediante cuestiones sencillas como «¿cuál es la frecuencia relativa de los casos que corresponden a la categoría C y cuya variable xi supera el valor Xi?». Para ello existen incluso lenguajes computacionales «ad hoc» (como el «Structured Query Language», SQL, para bases de datos relacionales). Pero, también como en el caso de la visualización, un empleo más inteligente de las preguntas propicia la obtención de información valiosa, sobre todo si se dirigen a relaciones que incluyan variables intermedias, resultados y errores.

Limitaciones:

- dependencia de la habilidad del usuario
- puede producir equívocos

Ventajas:

- interpretabilidad
- abre vías creativas

Recomendaciones de uso:

- como técnica exploratoria y combinable con otras

Métodos semianalíticos (estadísticos)

Los métodos analíticos (estadísticos) más elaborados corresponden a la denominada como formulación bayesiana: se suponen conocidas las propiedades del problema (probabilidades «a priori» de las clases y verosimilitudes de las observaciones en clasificación; en estimación, densidades de probabilidad de la magnitud que se pretende estimar y verosimilitudes de los datos dada ésta), que constituyen un modelo, y se define una política de costes asociada a los errores; tras de lo que se minimiza analíticamente el coste medio de la clasificación o estimación que se va a realizar, encontrando así la solución al problema. Las dificultades se centran en la validez del conocimiento que se supone: si se adopta un modelo equivocado, estos métodos, teóricamente óptimos, se degradan muy sensiblemente.

La vía semianalítica consiste en extraer la información precisa para la formulación (verosimilitudes, etc.) a partir de los datos: con ello, se reduce el riesgo de adoptar un modelo desafortunado; especialmente si se recurre para ello a las llamadas técnicas no paramétricas (frecuencia relativa, vecinos más próximos, ventanas de Parzen, etc.), que no hacen hipótesis alguna sobre la forma de las componentes del modelo.

En estadística clásica se recurre también a los métodos frecuentistas, que se basan en los estimadores muestrales para la estimación de los estadísticos de las variables: un estimador muestral simplemente promedia valores observados de la magnitud de que se trate. En el caso de la clasificación, se aceptan para las hipótesis modelos sencillos que difieren en sus parámetros, y se procede contrastando los efectos que producen las estimaciones de éstos por vía

muestral. Se trata, *per se,* de vías de tipo semianalítico: cuya potencia queda limitada por los modelos que se adopten.

Limitaciones:

- carácter subóptimo
- robustez limitada

Ventajas:

- potencia intermedia-alta
- hay costumbre para interpretar sus resultados

Recomendaciones de uso:

- en problemas sencillos o de los que haya un conocimiento razonable

Árboles de decisión

Son modelos lógicos que representan una partición binaria recursiva del conjunto de muestras etiquetadas disponibles atendiendo a decisiones o preguntas sobre una o más variables, obteniéndose finalmente un modelo de decisión en forma de árbol binario, lo que les da su nombre.
En la terminología de árboles se maneja una serie de elementos, que comprenderemos mejor haciendo referencia a un caso particular de toma de decisión acerca de la realización de una operación financiera mediante crédito o contado:

- *nodo raíz:* primer nodo del árbol, que representa el conjunto de todos los datos antes de realizar ninguna partición.
- *nodo hoja:* nodo terminal del árbol, que representa uno de los subconjuntos de datos que tienen asociada determinada decisión final.

- *atributos:* variables de las que depende tomar una decisión u otra, en este ejemplo, los atributos son «importe» y «solvencia».
- *tests o reglas:* evaluaciones de los atributos que deciden si se toma una decisión final o bien se prosigue con la evaluación.
- *decisión final:* en este caso es binaria: se decide entre «contado» o «crédito».

Para un patrón de entrada dado, si se recorre el árbol desde el nodo raíz hasta una de las hojas finales, se obtendrá la decisión correspondiente para dicho patrón. En la siguiente figura representamos el árbol de decisión resultante:

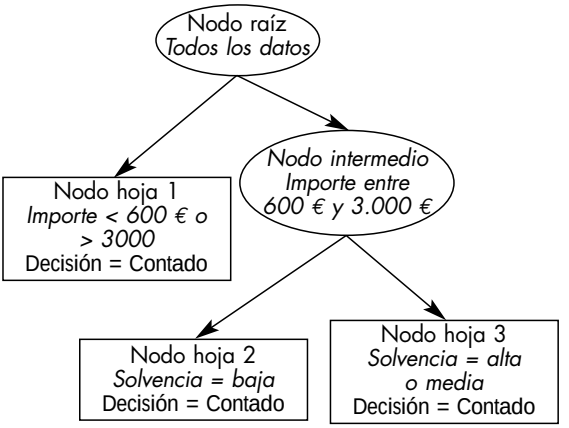

Fig. AI.2

Arbol de decisión binaria

La interpretación de dicha figura es inmediata, y se resume con los siguientes predicados:

- Si el importe es menor de 600 €, entonces la decisión = contado.

- Si el importe es mayor de 3.000 €, entonces la decisión = contado.
- Si el importe está entre 600 € y 3.000 € y la solvencia es alta, entonces la decisión = crédito.
- Si el importe está entre 600 € y 3.000 € y la solvencia es media, entonces la decisión = crédito.
- Si el importe está entre 600 € y 3.000 € y la solvencia es baja, entonces la decisión = contado.

Existen métodos inductivos muy eficientes que, a partir de un conjunto de ejemplos etiquetados, permiten estimar el conjunto de tests o condiciones que se deban aplicar, así como la estructura en árbol resultante. De entre los más conocidos podemos citar los algoritmos ID3 y C4.5.

Limitaciones:

- construcción no inmediata
- mala generalización
- adaptatividad dificultosa

Ventajas:

- fácil e intuitiva interpretación

Recomendaciones de uso:

- en problemas generales, cuando no sea crítica la obtención de una solución (cercana a la) óptima.

Sistemas expertos (ES: «Expert Systems»)

Son esquemas lógicos que condensan el conocimiento de un experto humano sobre una determinada materia o ámbito de aplicación, de tal modo que permiten posterior-

mente imitar dicho comportamiento o razonamiento [10] experto.

Habitualmente, dicha explicitación del conocimiento humano se lleva a cabo mediante la definición, en una primera etapa, de categorías que representen a un determinado ámbito, por ejemplo: «comida», «bebida», «carne», «pescado», «vino tinto», «vino blanco». A continuación, el conocimiento experto se puede reflejar en conjuntos de reglas del tipo «si X entonces Y», ya que una de las principales aplicaciones de tales sistemas es la toma de decisiones. Por ejemplo, para el caso de las categorías anteriormente mencionadas, una regla procedente de un experto podría ser: «si eliges carne como comida, entonces elige vino tinto como bebida», o bien «si eliges pescado como comida, entonces elige vino blanco como bebida». No obstante, también es posible encontrarse con esquemas más complejos/flexibles como los denominados de lógica difusa («Fuzzy Logic»), que permiten trabajar con decisiones «blandas» o grados de pertenencia de elementos a conjuntos («poco alto», «medianamente alto», «muy alto», o sus grados intermedios), o los que incorporan cierto componente de interpretación/análisis sintáctico-semántico de los datos presentados (en este caso, habitualmente en forma de lenguaje natural). Por continuar con el ejemplo anterior, en el caso de que la comida elegida sea un tipo particular de pescado de sabor fuerte, como el atún, la pertenencia del mismo a la clase «pescado» podría relajarse, de modo que se podría acercar un poco el atún al concepto «carne», hasta el punto que la regla de elegir vino blanco perdería fuerza frente a la que sugiere elegir vino tinto. Si a esta reponderación de los pesos de perte-

---

[10] Los términos «razonamiento» o «inteligencia» utilizados en esta sección, deben interpretarse con toda cautela y entenderse como mecanismos rudimentarios muchas veces simplemente utilizando lenguajes lógico-matemáticos. Todavía estamos lejos de ver auténticas máquinas «pensantes», si alguna vez son posibles…

nencia a las clases, añadimos alguna información a priori del tipo «el cliente Z prefiere casi siempre vino tinto», entonces el sistema casi con seguridad sugerirá tomar vino tinto con el atún, pese a ser éste un pescado. Estos últimos casos estarían plenamente vinculados a técnicas de «razonamiento» o inteligencia artificial («artificial inteligence», AI). Una vez incorporado dicho conocimiento humano en el sistema, la forma de operar ante nuevos casos por resolver suele llevarse a cabo mediante una metodología de pregunta-respuesta al usuario, de modo que el sistema experto interactúa con el usuario hasta producir una decisión final.

Una importante funcionalidad que se espera de un sistema experto es poder expresar o explicar de forma comprensible para un usuario humano las vías de «razonamiento» (reglas heurísticas) que le han llevado a tomar la decisión final.

No obstante, pese a ser un concepto muy interesante y a priori sencillo, su implementación práctica no está exenta de problemas graves: verificación de consistencia de las reglas introducidas por los expertos, control de la buena generalización ante nuevos casos, explicabilidad/interpretación con complejidad reducida, etc.

Limitaciones:

- construcción difícil
- potencia limitada
- mala generalización
- dificultad de obtener diseños adaptativos

Ventajas:

- proveen la interpretación

Recomendaciones de uso:

- en problemas muy concretos de complejidad moderada
- para situaciones forenses (directamente o extraídas como aproximación a otras máquinas)

Métodos basados en memoria

Son técnicas de fundamento muy sencillo: se almacenan en una memoria muestras etiquetadas y, ante la aparición de un nuevo caso, se busca en la memoria la muestra almacenada cuyas variables toman valores más similares a las del caso a resolver, y se le asigna a dicho nuevo caso como clase la de la muestra seleccionada de este modo.

En puridad, estos procedimientos no son más que aplicaciones directas de la técnica no paramétrica conocida como la del «vecino más próximo» («Nearest Neighbour», NN), muy tradicional en otras aplicaciones de clasificación de base estadística. En los métodos basados en memoria se suelen aplicar medidas de similitud variadas y que muchas veces tienen en cuenta el conocimiento de quien realiza el proceso de minería. Ni que decir tiene que una adecuada elección de la medida de similitud tiene importancia decisiva en las prestaciones de estos procedimientos.

Dicho lo anterior, se deduce de inmediato cuál es el defecto fundamental y la más inmediata vía de mejora de estas técnicas. El defecto radica en lo que se conoce como «amplificación del ruido»: aun suponiendo que las medidas de similitud sean acertadas, cualquier valor atípico en una instancia etiquetada contribuye a que muestras nuevas tiendan a clasificarse mal. Utilizar la extensión conocida como «K vecinos más proximos», en que se clasifica según la etiqueta mayoritaria de entre las de los K ejemplos de referencia más similares, reduce los efectos de la amplificación del ruido y tiende, en general, a proporcionar mejores prestaciones.

Limitaciones:

- gran dependencia de los datos y código
- gran dependencia del criterio de similitud aplicado
- prestaciones limitadas (tipo «vecino más próximo»)
- interpretación discutible

Ventajas:

- manejo muy sencillo
  (se espera mejorar su potencia en un futuro)

Recomendaciones de uso:

- en problemas muy concretos y razonablemente conocidos, de los que haya datos abundantes

Redes neuronales (NN: «Neural Networks»)

Son arquitecturas paralelo de elementos no lineales sencillos dispuestos en capas; la figura AI.3 presenta un esque-

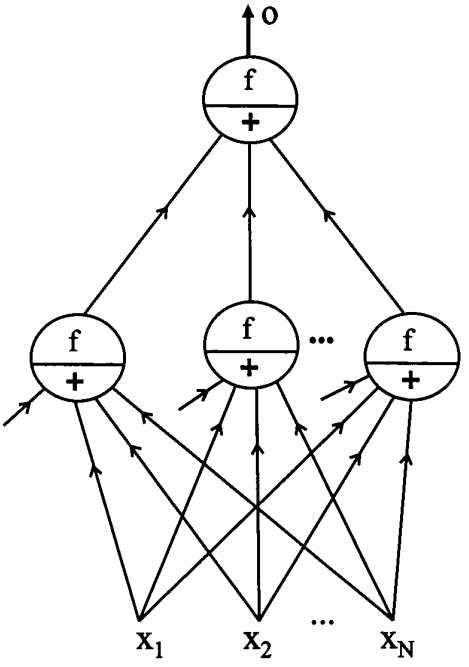

Fig. AI.3

Esquema de red neuronal tipo perceptrón multicapa

ma que corresponde a la arquitectura más conocida, el perceptrón multicapa («Multilayer Perceptron», MLP), en la que las flechas simbolizan multiplicaciones por los parámetros (w) de la máquina, y las no linealidades son típicamente tangentes hiperbólicas. No es ésta, se advierte, la arquitectura más recomendable, pese a su popularidad, para obtener las mejores prestaciones o para emplear en problemas complejos.
Para las redes neuronales en general:

Limitaciones:

- serias dificultades de diseño: elección de arquitectura, dimensionado, entrenamiento
- dificultad de interpretación

Ventajas:

- gran potencia
- posibilidad de buena generalización
- deterioro gradual ante fallos

Recomendaciones de uso:

- en situaciones generales: salvo necesidades forenses

TENDENCIAS

- En el diseño de sistemas expertos (y, en general, en el de todas las máquinas) va ganando aceptación el empleo de la «lógica difusa» («Fuzzy Logic»), en la que se sustituyen las dicotomías 0/1 por aproximaciones a probabilidades y se manejan las variables continuas como probabilidades de pertenecer a ciertos niveles cuánticos predefinidos sobre ellas (muy poco, poco, regular, mucho, muchísimo, p.ej.); aplicando unas reglas lógicas «difusas» que se dice que se asemejan a las que manejamos los humanos.

- En lo relativo a árboles, se trabaja en la automatización de la inserción de redes neuronales para las ramificaciones y, desde muy recientemente, en la concepción de versiones con ramificaciones «blandas» (es decir, no radicalmente excluyentes).
- Los métodos basados en memoria avanzan a partir de la proposición y prueba experimental de nuevos criterios de similitud, muchos de ellos «ad hoc»; y también se benefician de extensiones de su forma de clasificación (o estimación) basadas en la teoría de las «K vecinos más próximos» (o sus versiones interpolativas): lo que está aumentando sensiblemente su potencia y sus posibilidades generales.
- En el ámbito de las redes neuronales, la familia conocida como máquinas de vectores soporte («Support Vector Machines», SVM) ha demostrado impresionantes prestaciones, en particular para decisión dicotómica. Su fundamento (manejo de los datos como centros de «núcleos» no lineales a los que acceden las muestras que se han de tratar, y selección y ponderación de los mismos maximizando una medida de separación de las poblaciones de ejemplos) se está aprovechando para trazar muy numerosas y prometedoras vías para extenderlas (métodos de núcleos: «Kernel Methods»).
- También sobre todo en el ámbito de las redes neuronales (aunque no tan sólo en él) se desarrolla aceleradamente la tendencia a utilizar conjuntos, persiguiendo mejorar prestaciones en problemas de gran complejidad mediante el recurso a máquinas diferentes que pretenden resolver todo el problema (comités), o a máquinas análogas (o eventualmente distintas) que se concentran en diversos aspectos del problema (modulares) o que se ayudan para resolver conjuntamente el problema («boosting»). Debe resaltarse que de este modo se continúa en la línea «metafórica» inicial de las redes neuronales: el remedo de un sistema

nervioso, en el que, ciertamente, hay subsistemas cooperativos y especializados. Los conjuntos de conjuntos y sus planteamientos de diseño constituyen una línea de investigación particularmente prometedora; a la vez que todo este entramado apunta mejores posibilidades de interpretar la actuación de las redes neuronales.

- Finalmente, ha de decirse que no son desdeñables los esfuerzos que se están dedicando a la concepción de sistemas híbridos (conjuntos de máquinas de diferentes tipos, buscando el aprovechamiento de sus fortalezas) y mixtos (en los que las máquinas se destinan a mejorar diseños semianalíticos).

TIPOS DE HERRAMIENTAS

Para completar este apéndice dedicaremos unas líneas a discutir comparativamente los tres tipos básicos de herramientas que incorporan tecnologías de las citadas:

- Comerciales específicas (véase una lista seleccionada en el apéndice correspondiente).
- Comerciales generales: como puede ser el NN Toolbox de Matlab® (Mathworks) para las redes neuronales.
- Diseños «ad hoc».

si bien sus ventajas e inconvenientes están claros:

- Las herramientas comerciales específicas son de uso relativamente fácil e incorporan elementos para llevar a cabo todo el proceso de minería; a cambio, son rígidas y raramente permiten obtener muy buenas prestaciones, al no incluir avances recientes.
- Los diseños «ad hoc» se encuentran justamente en el extremo contrario.
- Y un lugar intermedio es el que ocupan las herramientas comerciales generales; raramente se hace un razo-

nable uso de las mismas a la vista de ello; de modo que conviene recordar que:

- Las herramientas comerciales generales resultan apropiadas para llevar a cabo primeros ensayos, normalmente exploratorios, sobre un problema, ya que permiten una cómoda y rápida apreciación de sus características, y pueden orientar en lo relativo a elecciones básicas (pretratamiento, tipo de máquina, etc.).
- En casos de datos masivos, conviene hacer uso de una herramienta comercial específica, por razones de completitud, comodidad, etc.; pero, si se desea elevar las prestaciones, no debe descartarse la inserción de módulos diseñados «ad hoc» para aquellas funciones en que resulte indicado.
- En el caso de volúmenes moderados de datos, cabe también la integración de módulos comerciales generales y, en su caso, «ad hoc».

APÉNDICE II
EJEMPLO ILUSTRATIVO DEL PROCESO DE LA MINERÍA DE DATOS

Presentamos aquí un caso imaginario de un supermercado que vende una serie de productos y desea conocer mejor a sus clientes para poder tomar así acciones de márquetin dirigido (selección de usuarios a los que enviar ofertas promocionales de nuevos productos). Todos los datos y situaciones presentados en este apéndice son ficticios, cuyo único fin es el de ilustrar un posible proceso de minería de datos para este escenario. El proceso de minería de datos lo simplificamos en los siguientes puntos:

- Uso de conocimiento experto sobre el negocio para determinar direcciones iniciales a seguir: qué pregunta hay que responder y qué dirección inicial se debe tomar.
- Identificación de datos iniciales y definición de su captura.
- Aplicación de alguna técnica de minería de datos que facilite el análisis de los datos.
- Interpretación experta y toma de nuevas decisiones antes de iniciar de nuevo el proceso.

En nuestro caso la pregunta está clara: ¿qué grupos de usuarios puedo identificar y caracterizar a fin de enviarles de forma selectiva ofertas promocionales de nuevos productos? Para ello, en primer lugar, utilizando conocimiento experto se

decide capturar datos de 12 usuarios midiendo una única variable: el número de cervezas adquiridas en la última semana. El resultado de dicha medida se resume en la tabla AII.1:

Usuario	N.º de cervezas
1	11
2	0
3	0
4	9
5	14
6	11
7	0
8	13
9	10
10	10
11	13
12	0

Tabla AII.1

Medidas tomadas de los 12 usuarios indicando el número de cervezas adquiridas en la última semana.

De la simple observación de la misma se puede deducir que existen dos grupos de usuarios claramente diferenciados: los que han comprado cerveza y los que no; pero esta información no resulta suficiente ni relevante para identificar o caracterizar dichos grupos, por lo cual se recurre a una toma de datos complementaria, completando la información anterior con la medida de una nueva [11] va-

[11] El proceso de identificación de nuevas variables relevantes en un problema dado es una tarea complicada y, aunque aquí se presenta como resuelta, es muchas veces el propio proceso de minería de datos es el que tiene que responder a dicha pregunta.

riable, en este caso, el número de pañales adquirido por cada usuario. Los resultados son los indicados en la siguiente tabla:

Usuario	N.° de cervezas	N.° de pañales
1	11	8
2	0	10
3	0	14
4	9	12
5	14	0
6	11	0
7	0	12
8	13	0
9	10	0
10	10	9
11	13	11
12	0	13

Tabla AII.2

Medidas tomadas de los 12 usuarios indicando el número de cervezas y pañales adquiridos en la última semana.

En este caso ya no es tan sencillo o inmediato obtener una interpretación, ya que la manipulación directa de números no es tarea natural para el ser humano. Se recurre por ello a una técnica de visualización, en la que se representa a cada usuario mediante un círculo en una gráfica bidimensional (N.° de cervezas en el eje X, N.° de pañales en el eje Y), tal y como se puede observar en la figura AII.1.

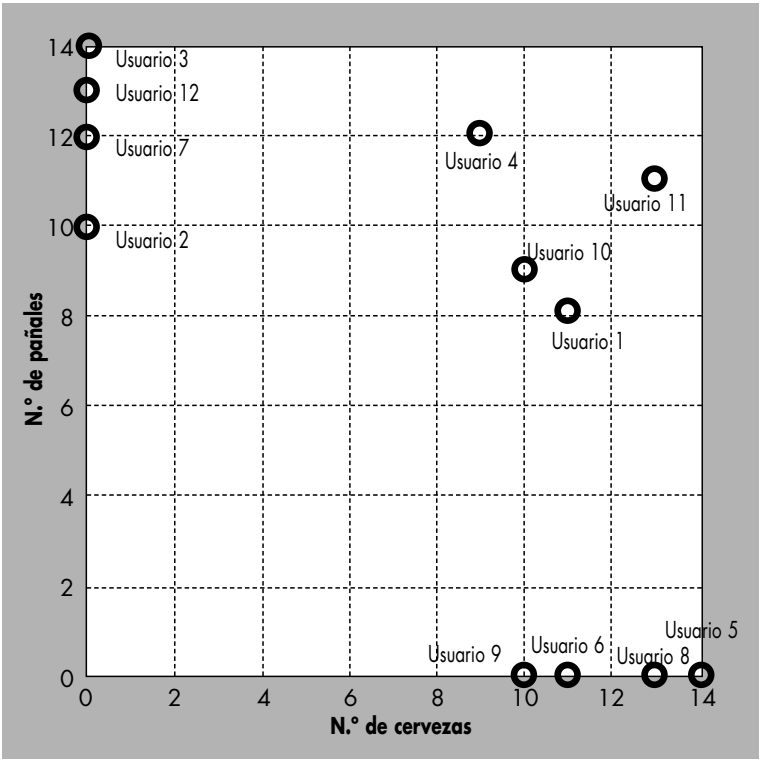

Figura AII.1

Visualización de los datos en función de las dos variables elegidas: N° de cervezas y N.° de pañales, cada círculo representa las medidas de un usuario con respecto al número de unidades compradas.

Con esta nueva representación es posible identificar claramente tres grupos de usuarios, tal y como se detalla en la figura AII.2.

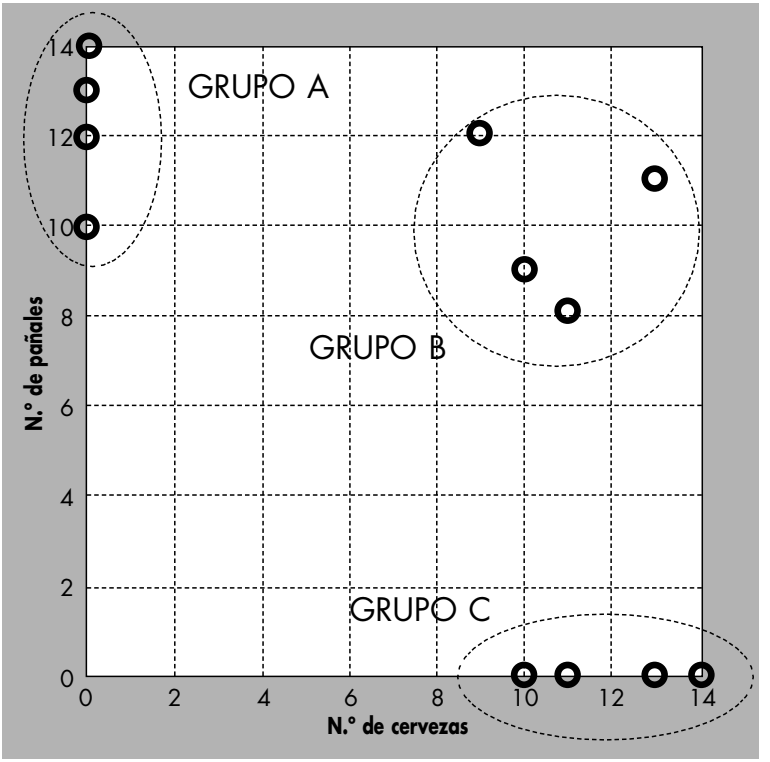

Figura AII.2

Visualización de los datos en función de las dos variables elegidas: N.° de cervezas y N.° de pañales, cada círculo representa las medidas de un usuario con respecto al número de unidades compradas.

Podemos proceder ahora a la interpretación de los tres grupos identificados:

- *Grupo A:* Denominaremos a este grupo como «madres», pues su perfil, atendiendo a las variables seleccionadas, se centra en la compra de pañales.
- *Grupo B:* Se adquieren simultáneamente cervezas y pañales, denominaremos a este grupo «padres».
- *Grupo C:* La compra se centra en cervezas, denominaremos a estos usuarios «solteros».

Atendiendo a esta interpretación de los resultados, es posible planificar una campaña de márquetin dirigido, mediante el envío de muestras de productos a los usuarios seleccionados. Así, se decide promocionar tres nuevos productos:

- *Una crema hidratante para bebés:* se envía una muestra del producto a los usuarios del Grupo A.
- *Una nueva revista de motor:* se envía una muestra del producto a los usuarios del Grupo B.
- *Una promoción de pizzas congeladas:* se envía una muestra del producto a los usuarios del Grupo C.

Es importante realizar de modo selectivo estos envíos, pues suponen un coste y es siempre interesante maximizar los retornos, esto es, pretender y lograr que el mayor porcentaje posible de usuarios que recibe una muestra, finalmente adquiera el producto de forma sistemática. A fin de evaluar la efectividad del análisis de datos llevado a cabo, se realiza una prueba piloto en la que se eligen, además de los tres grupos indicados, otros tres grupos de control, cuyos usuarios son elegidos al azar, a los que también se les envían las muestras. Se calcula el retorno (porcentaje de usuarios de cada grupo que efectivamente ha comprado con posterioridad el producto). Los resultados se han resumido como sigue:

	Grupo seleccionado	**Grupo aleatorio de control**
Crema	60%	12%
Revista	50%	10%
Pizza	70%	20%

Tabla AII.3

Porcentaje de retorno (éxito en la venta del producto) para cada una de los productos a en promoción y cada uno de los grupos de usuarios.

Como se puede observar, el porcentaje de éxito es mucho mayor dentro de los grupos de usuarios identificados a través del proceso de minería, con lo cual se puede concluir que el análisis de datos llevado a cabo ha sido efectivo y ha repercutido de forma positiva en la economía de la empresa.

En casos más generales, pueden ser miles las variables utilizables y cientos de miles los datos de usuarios que se pueden procesar, lo que hace muchas veces inviables los procesos de visualización directa, requiriéndose su sustitución por técnicas automáticas de agrupamiento, selección de variables, decisión máquina u otras muchas de las mencionadas en este documento.

Hemos visto mediante este ejemplo como, aun con métodos simples, ya se pueden iniciar procesos de minería: partiendo de conocimiento experto inicial del problema que hay que resolver, definiendo los primeros datos que se han de procesar y realizando operaciones relativamente sencillas, cerrándose un primer círculo de proceso mediante la interpretación de resultados. Una vez concluido un primer análisis, se pueden incorporar los resultados al propio proceso de minería, decidiendo qué nuevas variables incorporar, qué nuevas herramientas utilizar, qué nuevas preguntas formular, etc. Por ello, es importante resaltar que la implantación de soluciones de minería de datos es algo que está al alcance de todo el mundo, no siendo exclusiva su aplicación en grandes corporaciones mediante el uso de sofisticadas herramientas, pues en general es posible su implantación gradual, comenzando por pequeños retos que se resuelven mediante herramientas estándar (Excel, Matlab, visualizadores de datos, etc.), para ir progresivamente ampliando las exigencias, lo que suele llevar aparejado la necesidad de nuevas herramientas más completas o específicas.

APÉNDICE III
SOLUCIÓN AL TEST PERCEPTUAL DE LA FIGURA 1

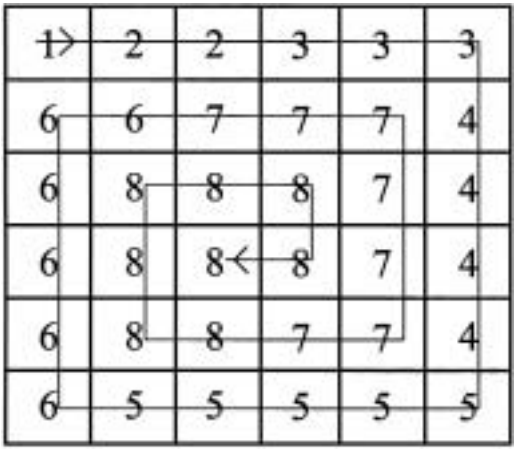

Figura AIII.1

Recorriendo el tablero según la espiral indicada se obtiene la secuencia 1 (círculo), 2 (cuadrados), 3 (triángulos), 4 (exágonos), 5 (cuadrados), 6 (círculos), 7 (triángulos) y 8 (exágonos).

GLOSARIO

Nos limitamos a algunos acrónimos (versión inglesa) que, además de aparecer en este documento, son frecuentes en la literatura sobre Minería de Datos y sus aledaños.

CRM («Customer Relationship Management»): Gestión de la relación con el cliente (individualizada, considerando sus necesidades y preferencias).

DM («Data Mining»): Minería de datos (proceso de extracción de información relevante de una base de datos).

DW («Data Warehouse»): Almacén de datos (en una gran base preparada para su fácil gestión a los efectos de acceder y procesar los datos).

ES («Expert System»): Sistema experto (conjunto de reglas para la resolución de un problema).

KD («Knowledge Discovery»): Descubrimiento de conocimiento (en fuentes en donde haya información o datos relevantes).

KM(1) («Knowledge Management»): Gestión del conocimiento (típicamente en una organización, de la información interna y externa y de la derivable de sus recursos humanos).

KM(2) («Kernel Methods»): Métodos de núcleos (forma particularmente potente de decisores y estimadores, emparentada con las redes neuronales y también con los métodos estadísticos clásicos).

MLP («Multilayer Perceptron»): Perceptrón multicapa (forma particular de red neuronal).

NN («Neural Networks»): Redes neuronales (arquitecturas algorítmicas paralelo con parámetros entrenables para resolver problemas de decisión, estimación, etc.).

OLAP («On Line Analytical Processing»): Procesado analítico «On Line» (software para manejar estadísticamente datos en tiempo real).

SQL («Standard Query Language»): Lenguaje estándar de preguntas (que se formulan sobre el contenido de una base de datos).

SVM («Support Vector Machines»): Máquinas de vectores soporte (arquitecturas algorítmicas construidas en función de las muestras más relevantes para resolver un problema, incluyendo arquitecturas de tipo neuronal).

TM («Text Mining»): Minería de textos (proceso de extracción de información relevante a partir de textos mediante algoritmos computacionales).

WM («Web Mining»): Minería en la web (minería de datos aplicada sobre la «World Wide Web»).

DOCUMENTOS COTEC sobre OPORTUNIDADES TECNOLÓGICAS

Documentos editados

N.º 1: Sensores.

N.º 2: Servicios de información técnica.

N.º 3: Simulación.

N.º 4: Propiedad industrial.

N.º 5: Soluciones microelectrónicas (ASICs) para todos los sectores industriales.

N.º 6: Tuberías de polietileno para conducción de agua potable.

N.º 7: Actividades turísticas.

N.º 8: Las PYMES y las telecomunicaciones.

N.º 9: Química verde.

N.º 10: Biotecnología.

N.º 11: Informática en la Pequeña y Mediana Empresa.

N.º 12: La telemática en el sector de transporte.

N.º 13: Redes neuronales.

N.º 14: Vigilancia tecnológica.

N.º 15: Materiales innovadores. Superconductores y materiales de recubrimiento.

N.º 16: Productos alimentarios intermedios (PAI).

N.º 17: Aspectos jurídicos de la gestión de la innovación.

N.º 18: Comercio y negocios en la sociedad de la información.

N.º 19: Materiales magnéticos.

N.º 20: Los incentivos fiscales a la innovación.

N.º 21: Minería de datos.

DOCUMENTOS COTEC sobre NECESIDADES TECNOLÓGICAS

Documentos editados:

N.° 1: Sector lácteo.

N.° 2: Rocas ornamentales.

N.° 3: Materiales de automoción.

N.° 4: Subsector agroindustrial de origen vegetal.

N.° 5: Industria frigorífica y medio ambiente.

N.° 6: Nuevos productos cárnicos con bajo contenido en grasa.

N.° 7: Productos pesqueros reestructurados.

N.° 8: Sector de la construcción.

N.° 9: Sector de la rehabilitación.

N.° 10: Aguas residuales.

N.° 11: Acuicultura.

N.° 12: Reducción de emisiones atmosféricas industriales.

N.° 13: El mantenimiento como gestión de valor para la empresa.

N.° 14: Productos lácteos.

N.° 15: Conservas vegetales.