Proceso de minería de datos

PID_00284574

Julià Minguillón Alfonso
Ramon Caihuelas Quiles

Tiempo mínimo de dedicación recomendado: 4 horas

Universitat Oberta de Catalunya

Julià Minguillón Alfonso

Licenciado en Ingeniería Informática por la Universidad Autónoma de Barcelona (UAB) en 1995, máster de Combinatoria y Comunicación Digital por la UAB en 1997 y doctor ingeniero en Informática por la UAB en 2002. Desde 2001 ejerce como profesor en los Estudios de Informática, Multimedia y Telecomunicación de la Universitat Oberta de Catalunya (UOC). Pertenece al grupo de investigación Learning Analytics for Innovation and Knowledge Application (LAIKA), donde desarrolla proyectos de investigación relacionados con el análisis y la visualización del comportamiento de los usuarios de entornos virtuales de aprendizaje y redes sociales.

Ramon Caihuelas Quiles

Licenciado en Ciencias de la Información por la Universidad Autónoma de Barcelona (UAB). Postgrado de Diseño de Aplicaciones por la Universidad Politécnica de Catalunya (UPC). Máster de Gestión de Tecnologías de la Información por Ingeniería La Salle URL. Doctorando en Informática por Ingeniería La Salle URL. Actualmente trabaja como responsable del grupo de bases de datos y soporte al desarrollo del Área de Tecnologías de la Universidad de Barcelona y como colaborador docente de los estudios de Informática de Ingeniería La Salle URL y la Universitat Oberta de Catalunya (UOC).

El encargo y la creación de este recurso de aprendizaje UOC han sido coordinados por el profesor: Julià Minguillón Alfonso

Primera edición: septiembre 2021

Av. Tibidabo, 39-43, 08035 Barcelona
Autoría: Julià Minguillón Alfonso, Ramon Caihuelas Quiles
Producción: FUOC

Índice general

Introducción

La minería de datos tiene que inscribirse dentro de un proceso de alcance más amplio: el descubrimiento de conocimiento dentro de grandes bases de datos, o KDD. KDD es la sigla del término inglés *knowledge discovery in databases*.

Nota

Este módulo es una revisión del módulo previo escrito por Ramon Sangüesa i Solé.

En este módulo definimos los objetivos del proceso de descubrimiento de conocimiento a partir de datos, delimitamos sus distintas fases, nos centramos en la fase de minería de datos y comentamos los problemas más frecuentes e importantes inherentes a todo el proceso.

Objetivos

Los objetivos de este módulo son los siguientes:

1. Tener una idea clara de todas las fases que comporta un proyecto de minería de datos.
2. Conocer la razón de ser de cada una de las fases del proyecto.
3. Anticiparse a los problemas concretos que se presentan en cada una de las fases del proyecto.

1. Descubrimiento de conocimiento en grandes volúmenes de datos

Desde un punto de vista académico, se considera la minería de datos como parte de un proceso mayor llamado *descubrimiento de conocimiento a partir de datos*. Sin embargo, actualmente algunos autores utilizan indistintamente las dos denominaciones (*descubrimiento de conocimiento* y *minería de datos*). Nosotros también usaremos los dos términos de manera indiferente.

Así, en cuanto al concepto de *descubrimiento de conocimiento* en grandes bases de datos (KDD), Piatetsky-Shapiro propuso la definición siguiente:

«El proceso de KDD (knowledge discovery in databases) es el proceso no trivial consistente en descubrir patrones válidos, nuevos, potencialmente útiles y comprensibles dentro de un conjunto de datos».

Referencia bibliográfica

El artículo original donde aparece esta definición es: G. Piatetsky-Shapiro; C. Mateus; P. Smyth; R. Uthurusamy (1993). «KDD-93: Progress and Challenges in Knowledge Discovery in Databases». *AI Magazine* (vol. 15, núm. 3, págs. 77-87).

Remarcamos aquí algunos aspectos que pueden pasar por alto en una primera lectura de la definición que acabamos de apuntar:

1) «Patrones válidos»: hay que entender este concepto como conocimiento correcto contrastable con la realidad.

2) «Potencialmente útiles»: la utilidad se encuentra en relación con el objetivo que nos proponemos al llevar a cabo el proceso de minería de datos.

Ejemplo de utilidad potencial

El hecho de saber que los clientes con rentas altas compran aparatos electrónicos puede ser útil si queremos distinguir los patrones de compra para cada nivel de renta, aunque es completamente inútil para decidir si un cliente será fiel a la empresa durante los próximos seis meses.

3) «Comprensibles»: la comprensibilidad está relacionada con el usuario que maneja el conocimiento, los patrones, etc., extraídos de los datos. De manera que no es una propiedad absoluta de los patrones obtenidos.

Ejemplo de comprensibilidad

Pongamos por caso que queremos predecir la distribución geográfica de ventas. Para un estadístico, los parámetros de un modelo de regresión es lo suficientemente comprensible; para un usuario menos preparado, quizá sea más comprensible ver un gráfico en pantalla.

Otra definición que pone énfasis en la parte que corresponde a la minería de datos es la que presentaron Holsheimer y Siebes. Dichos autores definen la minería de datos de la manera siguiente:

«La minería de datos es el proceso consistente en encontrar modelos comprensibles a partir de grandes volúmenes de datos».

En esta definición lo realmente importante es la palabra *modelo*. Un modelo es una descripción articulada y abstracta de una realidad.

Referencia bibliográfica

El trabajo original donde aparece esta definición es: M. Holsheimer; A. Siebes (1994). *Data mining: the Search for Knowledge in Databases*. Reporte técnico CS-R9406 (enero). Ámsterdam: Centrum voor Winskunde en Informatica (CWI).

Para describir un edificio, podemos tener un modelo que se base únicamente en conceptos estructurales, o que solo tenga en cuenta la distribución de conductos eléctricos. Lo que sí es importante y debemos tener en cuenta es que hay unos modelos que se avienen mejor que otros al tipo de utilidad que queremos obtener de los datos. Ya volveremos a tratar este tema más adelante.

La extracción del conocimiento a partir de bases de datos es un proceso complejo que se dirige a la obtención de modelos de conocimiento a partir de datos recogidos en una o más bases de datos en virtud de unos objetivos determinados. Este proceso consistente en pasar de datos a conocimiento supone un cambio de lenguaje de expresión e implica varias fases.

Los objetivos del proyecto de minería de datos determinan el tipo de modelo de conocimiento que tenemos que extraer. También determinan que un dato o una relación entre datos sea significativa o no en relación con los objetivos que nos hemos propuesto.

El cambio de lenguaje de expresión es importante, y aquí es donde reside la potencia del descubrimiento de bases de datos. Así, lo que hace un proceso de este tipo es extraer, a partir de los datos, un resumen en un lenguaje diferente, pero:

- Comprensible para quien lleva a cabo el proyecto cuando analiza el conocimiento extraído.

- Directamente integrable en las operaciones de la empresa, aunque no siempre lo es.

Ejemplo de cambio de lenguaje de expresión

Si partimos de un conjunto de registros de bases de datos extraídos a partir de los códigos de barras del terminal de punto de venta, tendremos una repetición de valores de productos (por ejemplo, margarina) y códigos postales (por ejemplo, 08012). Un resumen que solo extraiga esta ocurrencia representa un cambio de lenguaje poco potente. En cambio, uno que nos dé como resultado una regla del tipo:

Si código postal = 08012, entonces compra(margarina)

ha efectuado un cambio hacia un lenguaje más comprensible. El conocimiento de que los clientes que compran margarina viven en la periferia de la ciudad y los que compran

mantequilla, en el centro, puede ser una relación significativa o no dependiendo del objetivo que nos hayamos propuesto para llevar a cabo el proceso de descubrimiento.

1.1. ¿Qué entendemos por *conocimiento*?

No entraremos aquí discusiones muy profundas acerca de qué es el conocimiento. A efectos prácticos, podemos asimilar el concepto *conocimiento* a una información que ha sido interpretada, clasificada, aplicada y revisada, de manera que tiene un cierto valor para el usuario de la información inicial en cuanto a sus objetivos. Si se quiere, podemos adherirnos al concepto de conocimiento como «creencia justificada», en el sentido de que interpretamos en relación con lo que sabemos y con el hecho de relacionar los datos con lo que sabemos. Así es como obtenemos conocimiento nuevo. Notad que en esta acepción, el conocimiento siempre se encuentra sometido a revisión a la luz de nuevas informaciones.

Podemos pensar en el conocimiento como creencia justificada. Podríamos tener la creencia intuitiva de que la mayoría de los clientes de una empresa son de Barcelona, pero si los datos nos dicen que 15,000 clientes son de Barcelona y 30,000 del resto de Cataluña, tendremos que interpretar el hecho como que la mayoría de los clientes de una empresa son de fuera de Barcelona. Nuestra creencia inicial no queda, pues, justificada, y no es un conocimiento válido. Las opiniones y los prejuicios suelen estar equivocados si no se basan en datos reales.

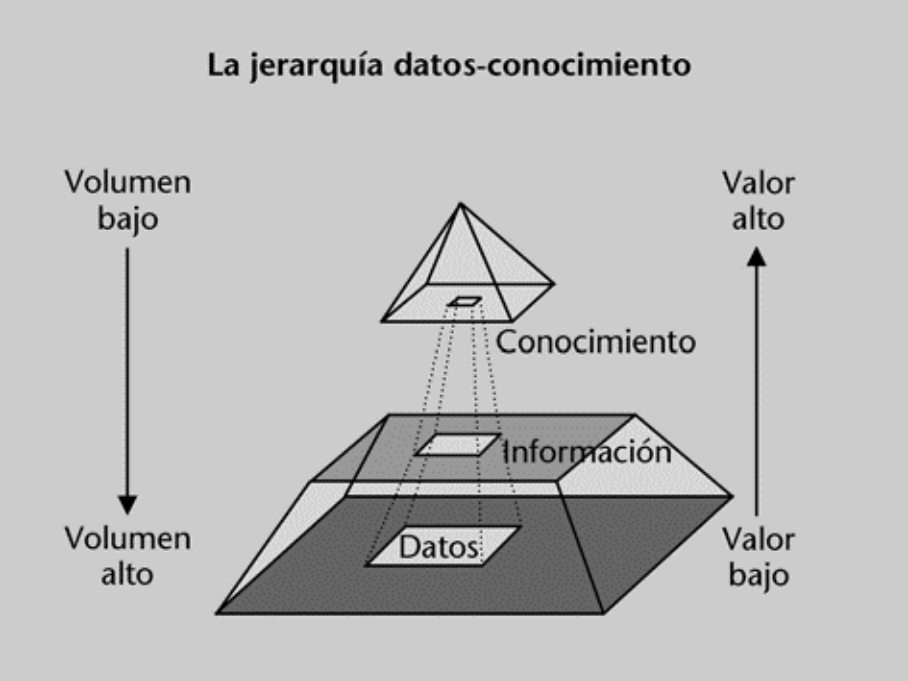
Solo podemos llegar a una justificación de nuestra creencia cuando relacionamos los datos en bruto (el número de clientes de cada zona) con otro dato procedente de nuestro conocimiento previo (por ejemplo, que tenemos 45,000 clientes). Haríamos otra interpretación si el número de clientes fuera 250,000: tendríamos que relativizar la creencia, ya contrastada, de que la mayoría de los clientes (30,000) eran de fuera de Barcelona.

Lo que se pide al proceso de minería de datos es que aporte un primer nivel de interpretación mediante la extracción de relaciones con datos en bruto. De este primer nivel se puede extraer conocimiento todavía más elaborado.

Existe consenso en considerar que en el proceso de extracción de conocimiento los datos son la materia prima; cuando alguien les atribuye un significado, tenemos información; cuando se hace una abstracción de esta información en relación con los conceptos necesarios y relacionados con un objetivo, tenemos un conocimiento.

La figura 1 refleja muy bien la jerarquía de importancia que se desarrolla en el proceso de extracción de conocimiento.

Figura 1. Pirámide del conocimiento



Ejemplo de proceso de extracción de conocimiento

Aquí tenemos parte de un conjunto de datos que corresponde a los clientes de un gimnasio. Los conceptos que utilizamos para expresar el conocimiento que queremos obtener son la renta y el tipo de actividad deportiva:

Tabla 1

Grupo	Centro	Horario	Act1	Act2	Renta	Edad	Sexo
1	1	Mañana	Yoga	*Stretch*	Alta	68	M
2	3	Tarde	Yoga	*Steps*	Media	32	H
4	3	Tarde	*Stretch*	Yoga	Baja	44	M
2	3	Tarde	*Steps*	Pesas	Media	23	H
1	3	Tarde	Pesas	*Stretch*	Media	35	M
2	1	Mañana	Pesas	Pesas	Media	45	M
2	1	Mañana	Yoga	*Steps*	Baja	19	M
1	2	Mañana	*Stretch*	*Stretch*	Alta	21	M
3	3	Mañana	*Steps*	Aeróbic	Alta	56	H
3	1	Mañana	Aeróbic	*Steps*	Baja	30	M

Un primer resumen de los datos nos dice que hay tres clientes de renta alta, cuatro, de renta media y tres, de renta baja. También nos dice que hay tres clientes que hacen como actividad principal (Act1) yoga, dos que hacen pesas, dos, estiramientos (*stretch*), uno, aeróbic y dos, *steps*. Ahora, aplicando técnicas tradicionales de estadística sobre todo el conjunto original –que contiene más registros que los que hemos presentado en esta tabla–, podemos decir que hay una correlación de 0.8 entre la renta y la actividad principal, lo cual ya es un primer cambio de lenguaje.

Aplicando alguna otra técnica de minería de datos podríamos extraer una lista de reglas de este tipo:

```
If Act1 is Steps Then
   Renta is Media
   Rule's probability: 0.981.
   The rule exists in 52 records.
   Significance Level: Error probability < 0.2.
```

que nos indica que si la actividad principal de un cliente es *steps*, entonces podemos asegurar con una probabilidad del 98.1 % que su renta es media, que ese aspecto ha sido observado en cincuenta y dos casos originales, y que la significación de esta regla es menor del 20 % (ahora no entramos en el significado de esto), lo cual es un nuevo cambio de lenguaje, e incluso más comprensible para alguien con conocimientos de estadística.

El cambio de lenguaje desde la materia prima hasta una expresión del conocimiento comprensible u operativa requiere varias fases. Las comentamos a continuación.

2. Las fases del proceso de extracción de conocimiento

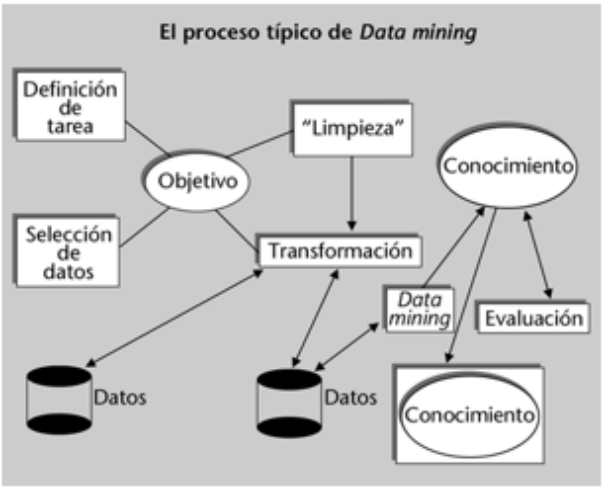
Como hemos dicho, la minería de datos no es un proceso que se realice en una sola fase. No es que decidamos expresar un tipo de objetivo determinado (predecir la continuidad de los clientes, por ejemplo) y automáticamente se genere un modelo que nos resuelva el objetivo (que nos diga si un nuevo cliente será fiel durante un cierto tiempo). Es más complicado y menos automático. Hay que pasar por varias fases.

Existen diferentes formulaciones de las fases para dividir el proceso. Seguiremos a Fayyad y esquematizaremos el proceso de descubrimiento según las fases que expresamos gráficamente en la figura 2.

Referencia bibliográfica

Encontraréis el artículo de las fases del proceso de extracción de conocimiento en la obra siguiente:
U. Fayyad; G. Piatetsky-Shapiro; P. Smyth (1996). «The KDD process for extracting useful knowledge from volumes of data». *Communications of the ACM* (vol. 39, núm. 11, págs. 27-34).

Figura 2. Proceso típico de *data mining*



Fuente: adaptado de Fayyad y otros (1996)

Estas fases son: definición de la tarea de minería de datos, selección de datos, preparación de datos, minería de datos propiamente dicha, evaluación e interpretación del modelo e integración. Como ya hemos señalado antes, este proceso no es lineal, sino que se realimenta y continúa: nuevos cambios en la situación pueden hacer que nuestro conocimiento deje de ser correcto, por lo que será preciso volver a extraer conocimiento nuevo.

2.1. Definición de la tarea de minería de datos

Este es el punto en el que precisamos cuál es el objetivo del proyecto de minería de datos. Así pues, es el momento de decidir si se trata, por ejemplo, de encontrar dependencias entre variables (la renta y la actividad, por ejem-

plo), si queremos saber qué distingue un tipo de usuario de otro, si queremos conocer tendencias o detectar patrones, etc.

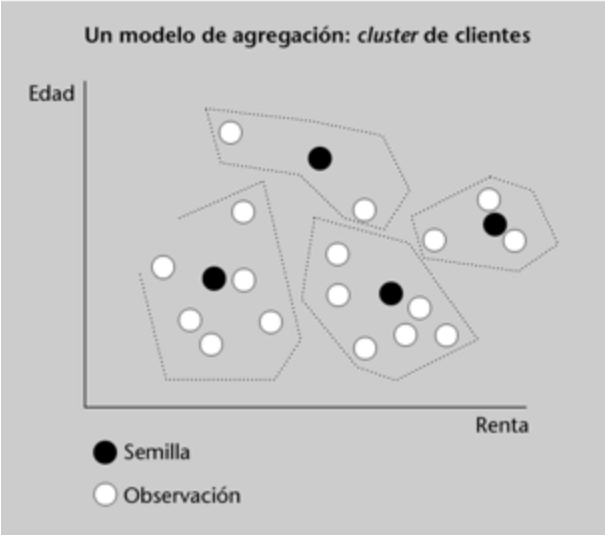
La tarea principal de cada uno de estos proyectos en general puede asimilarse a alguna de las siguientes:

1) Encontrar similitudes y agrupar objetos parecidos. Corresponde a un proyecto en el que tenemos «poca» información del dominio y queremos empezar a tener una idea más clara al respecto. Los modelos típicos para alcanzar estos objetivos son los modelos de agregación (*clustering*) procedentes del análisis de datos o del aprendizaje automático y los modelos asociativos.

Ejemplo de modelo de agregación

Un ejemplo típico de proyecto de minería de datos para encontrar similitudes consiste en encontrar grupos de clientes parecidos.

Figura 3. Modelo de agregación: *cluster* de clientes



2) Clasificar objetos. La tarea de estos proyectos de minería de datos no es exactamente igual a la del punto anterior. Aquí lo más habitual es partir de una situación más informada, sabiendo que existen grupos ya definidos. Lo que nos interesa en este caso es estudiar mejor las diferencias existentes entre un grupo y otro, sus características peculiares. Por norma general, la clasificación de objetos es el paso previo a la realización de predicciones; es decir, para saber alguna conducta de interés a partir de una serie de datos. Todo esto podemos refinarlo obteniendo conocimiento predictivo.

Lectura recomendada

Podéis encontrar más información sobre la evolución de estos modelos clasificatorios en el artículo siguiente:
X. Wu; V. Kumar; J. Ross Quinlan y otros (2008). «Top 10 algorithms in data mining». *Knowledge Information Systems* (vol. 14, págs. 1-37). https://doi.org/10.1007/s10115-007-0114-2

Algunos modelos clasificatorios típicos son los árboles de decisión como CART, ID3, C4.5 y C5.0; también se pueden mencionar las redes neuronales para clasificación y los sistemas basados en reglas de clasificación.

Los árboles de decisión ofrecen una estructura en la que en cada nodo se hace una pregunta sobre un atributo determinado. El valor que tome indica que hay que seguir la rama correspondiente al atributo. Los nodos finales corres-

ponden a conjuntos de ejemplos que pertenecen a la misma clase. Si seguimos las ramas desde la raíz hasta las hojas, se obtiene una serie de condiciones que permiten clasificar las nuevas observaciones. Por ejemplo, en la figura 4 podemos ver la estructura de un árbol de decisión cuyo objetivo es predecir si un cliente de un gimnasio solicitará los servicios de un entrenador personal o no.

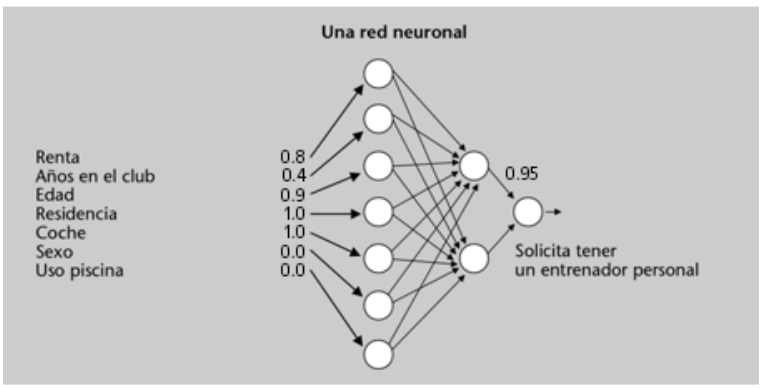
Figura 4. Árbol de decisión

Normalmente, los métodos de construcción de árboles de decisión vienen acompañados de más información que permite saber para cada nodo y sus posibles valores cómo queda dividido el conjunto de observaciones en varias clases.

Las redes neuronales también son buenos modelos clasificatorios y predictivos. Dichas redes presentan ciertas analogías con la manera en la que están conectadas las neuronas cerebrales y se organizan en forma de muchos nodos de proceso conectados que dan una o más salidas. Las distintas capas de nodos están conectadas entre sí con más o menos fuerza mediante unos factores o pesos que indican la importancia de las salidas que han tenido lugar por cada nodo. En conjunto, lo que hacen es aprender a ajustar los valores de estos pesos con el fin de ser lo más predictivas posible. En la actualidad, la capacidad de cómputo disponible ha hecho que resurjan con fuerza, con lo que han dado lugar a lo que se conoce como *deep learning*.

En la figura 5 podemos ver una red neuronal sencilla utilizada también para predecir si un cliente solicitará entrenador personal o no.

Figura 5. Red neuronal



Las entradas de la red recogen los valores de otros atributos que hay que tener en cuenta (sexo, edad, renta, residencia, etc.) y la salida tiene un valor cercano a 1 si, efectivamente, a la combinación de valores (descripción de un cliente) normalmente le corresponde un entrenador personal.

Por su parte, las reglas de clasificación tienen una expresión como esta:

$$\text{Antecedente} \Rightarrow \text{Consecuente}$$

Estas reglas imponen una serie de condiciones sobre los valores que toman los atributos de entrada con el fin de indicar a qué clase pueden pertenecer. En nuestro ejemplo, la clase estaba determinada mediante el atributo Entrenador personal. Hay dos clases, la de los clientes que solicitan entrenador y la de los que no. Así pues, la forma que tiene la regla nos indica bajo qué condiciones un cliente solicitaría tener entrenador:

```
If Act1 is Steps Then
   Entrenador personal is No
   Rule's probability: 0.981.
   The rule exists in 52 records.
   Significance Level: Error probability < 0.2.
```

Esta no es la única forma que admiten las reglas de clasificación, hay otras formas que pueden adoptar. Por ejemplo, las reglas de clasificación que se obtienen con métodos como el CN2 o las que se obtienen a partir de métodos de lógica inductiva, que son una conjunción de condiciones lógicas sobre los valores que pueden adoptar los atributos (igualdad, comparación, etc.) y no suelen tener ninguna indicación con respecto al grado de validez, significación o error. La regla análoga a la anterior, que puede obtenerse utilizando CN2, tiene la forma siguiente:

```
If (Act1 = Steps) => (Trainer = no)
```

3) Predecir. Se trata de obtener conocimiento que nos permita predecir aquello que nos interese. Presenta muchas similitudes con la clasificación. Efectivamente, en una clasificación binaria de dos clases posibles (clientes que solicitan entrenador y clientes que no lo solicitan), se trata de predecir el valor de la clase dentro de un conjunto limitado de valores (en este caso [0, 1], donde el 0 representa a quienes no piden entrenador y el 1, a quienes sí lo piden). En otros casos con más clases, la clasificación puede entenderse como un proceso de predicción que tiene que indicar el valor de esta etiqueta.

Ahora bien, en algunas ocasiones lo que nos interesa pedir no es un atributo que adopte valores en un conjunto finito de valores (numéricos o no), como

acabamos de ver. Por ejemplo, puede interesarnos predecir la duración de las llamadas que efectúa un departamento a partir de otros datos. En ese caso, estaremos intentando obtener un valor que varía en una escala continua de valores (incluso podríamos tener decimales); por lo tanto, el número de etiquetas de clase posibles es infinito: hay tantas como números reales entre el valor mínimo y el máximo detectados.

Existen muchos ejemplos de modelos predictivos: algunos ejemplos de modelos que nos permiten llegar a este objetivo son los árboles de decisión y los modelos predictivos clásicos de la estadística, por ejemplo, los modelos de regresión. Aquí podríamos incluir también los modelos que no detectan valores concretos (por ejemplo, sí o no, para la fidelidad de un cliente), sino que detectan tendencias. Típicamente, incluiríamos aquí los estudios de series temporales y todas las variaciones que se han aportado desde el aprendizaje automático, como las redes neuronales para la predicción de series temporales.

4) Describir. Aunque obtener agrupaciones de objetos es un primer nivel de descripción, con conocimiento descriptivo nos solemos referir a encontrar y expresar asociaciones significativas o causales entre diferentes variables.

Un ejemplo típico de este tipo de modelos es el de las redes bayesianas y, aunque con menor potencia, las reglas de asociación. Las presentamos brevemente a continuación.

Consideramos primero una red bayesiana de ejemplo, como la que vemos en la figura 6.

Figura 6. Red bayesiana

Esta red bayesiana indica cosas interesantes. Ha sido extraída de un conjunto de datos simplificados que relaciona varias variables utilizadas en seguros para predecir si un cliente es de riesgo o no. Tan solo inspeccionando visualmente el modelo, ya nos indica que las variables que influyen más directamente en la clase de riesgo de un cliente son el color del coche, el kilometraje, las multas y el año de matriculación. Asimismo, podemos ver que otras relaciones, como por ejemplo la edad del conductor o conductora, no influyen directamente en la clase, sino indirectamente por vía del color y en combinación con otras variables, como la renta del cliente. Además, para cada enlace podemos saber la probabilidad condicional existente entre las distintas variables conectadas. En tal caso, podemos ver los valores de la tabla 2.

Tabla 2. Tabla de probabilidades condicionales

Sexo	Edad	Kilometraje alto	Kilometraje bajo
Mujer	Joven	0.708	0.292
Mujer	Mayor	0.582	0.418
Hombre	Joven	0.714	0.296
Hombre	Mayor	0.255	0.645

Por lo tanto, tenemos una idea de la influencia mutua entre las variables, cuáles son realmente relevantes para conocer el valor de una variable dada y, además, cuando observamos una determinada combinación de valores, podemos predecir los valores más probables para el resto de las variables relacionadas.

Por su parte, las reglas de asociación tratan de encontrar coocurrencias lo suficientemente significativas entre grupos de variables. El único requisito que imponen es que se indique el «nivel de soporte» que se quiere que tengan a partir de los datos, la proporción de los datos que nos interesa cubrir con esta regla. Entonces, hay que encontrar grupos de variables y combinaciones de valores que lleguen a tener este grado de soporte. Por ejemplo, se da el caso siguiente:

```
(Renta > 80,000) & (Edad > 40) & (residencia = C) =>
(Entrenador = Si) & (Piscina = Si) (0.95)
```

En otros términos, en un 95 % de los casos, cuando el cliente tiene una renta que supera los 80,000 euros, una edad superior a los cuarenta años y vive en la zona que hemos designado como C, entonces también resulta que ha pedido entrenador personal y utiliza la piscina.

Ejemplo de modelo descriptivo

En el caso de los seguros, un modelo descriptivo pondrá de relieve que los factores determinantes de la peligrosidad de un cliente son su edad, el color del coche que compre, el año de matriculación y el valor de los atestados presentados en el último año. Además, nos indicará el tipo o nivel de fuerza de asociación existente entre todas estas variables. Los modelos descriptivos son muy fáciles de interpretar si se presentan en un lenguaje cercano al experto en el ámbito (en este caso, los seguros).

5) Explicar. Aquí se trata de obtener modelos que puedan darnos las razones de por qué se ha producido un comportamiento determinado. Por ejemplo, si habíamos predicho que un cliente determinado adquiriría una serie de productos determinados y no lo ha hecho, ¿a qué puede deberse? Ejemplos de modelos de este tipo son las redes bayesianas, que permiten determinar cuál es el conjunto de variables, y con qué valores para un valor observado por una variable determinada, que más probablemente pueden dar razón del valor observado.

Lectura recomendada

Encontraréis más información sobre los modelos explicativos en la obra siguiente:
T. Anand; G. Kahn (1992). «SPOTLIGHT: A Data Explanation System». *Proceedings of the Eighth IEEE Conference on Applied Artificial Intelligence* (págs. 2-8). Washington D.C.: IEEE Press.
Un trabajo más reciente sobre reglas es el siguiente:
M. Almutairi; F. Stahl; M. Bramer (2017). «Improving Modular Classification Rule Induction with G-Prism Using Dynamic Rule Term Boundaries». En: M. Bramer; M. Petridis (eds.). *Artificial Intelligence XXXIV*. SGAI 2017. Springer, Cham («Lecture Notes in Computer Science», 10630). https://doi.org/10.1007/978-3-319-71078-5_9

Veamos algunas consideraciones globales para todo tipo de tareas de minería de datos. Es importante que tengamos claro que para cada una de estas tareas se puede utilizar más de un modelo. Además, para construir cada tipo de modelo tenemos a nuestra disposición varios métodos que estarán incluidos, o no, en las herramientas comerciales existentes en el mercado, y cada una de estas puede avenirse, o no, con los condicionantes técnicos del sistema de base de datos del que disponemos y del sistema de información en el que se ubica. Por ejemplo, las redes bayesianas pueden ser utilizadas como modelos descriptivos, predictivos y explicativos.

Por lo tanto, a la hora de definir la tarea de minería de datos, hemos de ser capaces de:

- Aproximar el objetivo a alguna de las tareas genéricas que hemos mencionado. Este punto es el más crítico, y para el que resulta más difícil dar reglas de aplicabilidad general. Decidir, por ejemplo, que tenemos que conocer mejor a nuestros clientes es un objetivo demasiado general que necesita ser precisado. ¿Queremos conocer qué grupos de clientes tenemos? Entonces hemos de obtener un modelo de agregación. ¿Queremos determinar las características distintivas de grupos separados geográficamente o por renta? Entonces tenemos que clasificar. Sin embargo, sea como sea, siempre quedan objetivos que requieren un esfuerzo combinado. Quizá primero tendremos que agrupar, después clasificar y más tarde extraer un modelo predictivo.

- Decidir el modelo que necesitamos. Este punto nos obliga a conocer bien los distintos modelos que existen, para qué sirven y para qué no. Podemos compararlos, por ejemplo, en términos de su capacidad expresiva, de su comprensibilidad, de su facilidad de implementación e integración en el sistema de información de la empresa.

- Seleccionar el método necesario para construirlo. Este tercer punto está muy relacionado con el segundo y nos permitirá evaluar las posibilidades y las características de las distintas herramientas existentes: complejidad del método y coste computacional.

Los dos últimos puntos se encuentran, además, condicionados por el entorno donde tiene que desarrollarse el proyecto para la implementación de los resultados: sistema de información y entorno de explotación.

Una vez definido el objetivo, y cuando lo hemos puesto en relación con la tarea principal del proyecto, qué modelos nos interesan más y qué métodos y herramientas nos hacen falta, debemos pasar a encontrar la materia prima: los datos. No es tan sencillo como parece.

A continuación entramos a describir detalladamente cada fase, lo que nos permitirá abordar los módulos restantes con una perspectiva global.

Origen de los datos

Encontrar los datos que necesitamos es más fácil de decir que de hacer. Es el equivalente en la minería real a saber dónde tenemos que empezar a perforar para encontrar petróleo.

Desde una perspectiva ideal, la tecnología de *data warehousing* –literalmente, almacenes de datos– (Inmon, 1996) está especialmente orientada a facilitar la localización de los datos dentro de una empresa en relación con varios tipos de utilidades. Un *data warehouse* integra datos procedentes de los distintos datos de cada departamento de una empresa.

Con esta tecnología se asegura, por ejemplo, que para los distintos registros el campo nivel de riesgo del cliente tiene el mismo significado, aunque ha resultado de fusionar las diferentes interpretaciones que le da el departamento de cobros o el de pólizas –si hablamos, por ejemplo, de una compañía de seguros–. Además, los *data warehousing* guardan datos históricos de la empresa, de manera que permiten predecir tendencias. En principio, son una tecnología dirigida a la toma de decisiones. Entonces, ¿qué problema presentan? ¡Pues que la mayoría de las empresas no tienen esta tecnología instalada! Y esta afirmación se hace más cierta cuanto más pequeña es la empresa. En consecuencia, lo más normal es que, como primera fase, haya que localizar las fuentes de datos a partir de:

- Las bases de datos dispersas por los distintos departamentos que consideramos que pueden ser relevantes para el proyecto que llevamos a cabo.

- Las bases de datos transaccionales, que registran las operaciones día a día y que acumulan información histórica que puede ser relevante para obtener el modelo que estamos buscando.

Ninguna de estas dos operaciones es sencilla. De hecho, es necesario que la empresa –si no tiene un *data warehouse*– realice una buena política de gestión

de datos de cara a la toma de decisiones. Es muy probable que haya que crear e introducir nuevos procesos para obtener los datos que necesitamos.

Ejemplo sobre la necesidad de introducir nuevos procesos

Cuando una agencia de viajes decidió que quería recomendar mejor a sus clientes los posibles destinos de viaje en relación con qué viajes había hecho cada cliente anteriormente, se dio cuenta de que no guardaba información histórica útil de cada cliente más allá de un cierto número de meses, por lo que tuvo que introducir un nuevo proceso y modificar la base de datos correspondiente para disponer de datos anteriores en el tiempo, necesarios para poder extraer conocimiento más fiable y analizar tendencias y patrones estacionales.

2.2. Preparación de los datos

A continuación, comentaremos brevemente las técnicas que se utilizan para asegurar los tres aspectos que hemos mencionado antes. Son la limpieza de datos, la transformación de los datos y la reducción de la dimensionalidad.

Una vez localizadas las fuentes de datos, debemos proceder a prepararlas para que se les puedan aplicar los métodos o las herramientas que construirán al modelo deseado. Esta fase, aunque parezca sencilla, junto con la de selección de datos, consume el 70 % del esfuerzo (¡o más!) en los proyectos de minería de datos de nueva implantación.

En este punto, hay que asegurarse de unas cuantas cosas. Veámoslas:

1) Que los datos tengan la calidad suficiente. Es decir, que no contengan errores, redundancias o que presenten otros tipos de problemas. También se entiende la calidad de los datos como aquella propiedad que asegure la calidad del modelo resultante (por ejemplo, que asegure que será lo suficientemente predictivo, si se trata de crear un modelo de este tipo). Ni que decir tiene que esta última acepción de la palabra *calidad* es todavía más problemática de garantizar.

2) Que los datos sean los necesarios. Quizá los habrá que no nos hagan falta, y quizá tengamos que añadir otros. Suele ser muy extraño que los datos que necesitamos realmente ya hayan sido recogidos por el sistema con el propósito de llevar a cabo justamente el tipo de estudio de minería de datos que queremos emprender. Eso, normalmente, supondrá añadir campos nuevos a las distintas relaciones de una base de datos procedentes de otras relaciones o de otras bases de datos.

3) Que estén en la forma adecuada. Muchos métodos de construcción de modelos requieren que los datos estén en un formato determinado que no tiene que coincidir necesariamente con el formato en el que se encuentran almacenados. Comprobaremos que hay varias interpretaciones de esta diferencia de formato que obligan a efectuar diferentes transformaciones que estudiaremos en su momento. La más típica es que los datos sean valores numéricos continuos y los métodos solo admitan valores discretos.

Una práctica común es generar nuevas variables derivadas de expresiones más o menos complejas a partir de los datos originales. Esto permite originar indicadores que en sí mismos ya agregarán información y serán muy útiles en el proceso de conocer la información que guardan los datos y en el posterior proceso de creación del modelo.

Lectura recomendada

Encontraréis más información acerca de la fase de preparación de datos en la obra siguiente:
E. Simoudis; U. M. Fayyad (1997, marzo). *Data Mining Tutorial*. First International Conference on the Practical Applications for Knowledge Discovery in Data Bases. Londres.
Un trabajo más reciente es el siguiente:
S. García; J. Luengo; F. Herrera (2015). *Data preprocessing in data mining* (vol. 72). Cham, Suiza: Springer International Publishing.

2.2.1. Limpieza de los datos

La limpieza de datos consiste en procesar los datos para eliminar los que sean erróneos o redundantes. También hay quien incluye en esta fase el paso consistente en eliminar algunos de los atributos de los datos, lo que se conoce como selección de atributos, y cuya intención es asegurar que con menos datos se puedan obtener modelos de la misma calidad. Comentaremos este aspecto más adelante.

Aun estando en la forma adecuada, suele suceder que los datos no son perfectos en un 100 %. Los datos introducidos a mano o procedentes de la fusión de varias bases de datos suelen mostrar factores de distorsión importantes. Revisemos cuáles son los más frecuentes:

1) Datos incompletos. Puede suceder, especialmente en aquellos atributos en los que cuando se diseñó el proceso correspondiente de entrada de datos se decidió que no eran obligatorios o que tenían formato libre, que tengan un valor «indefinido», es decir, que falte algún valor de los que son comunes para los registros de la base de datos que estamos considerando. En inglés esto se conoce como *missing values*.

Normalmente, lo que se hace es «completar» los valores que no aparecen. Por ejemplo, para los valores numéricos se complementa calculando el valor medio observado para un atributo, aunque también es posible usar técnicas de imputación más sofisticadas.

Ejemplo de datos incompletos

Es el que tenemos cuando rellenamos el campo correspondiente a la calle, pero olvidamos, o no se representa correctamente, el correspondiente al número o al piso. Una manera de solucionar los datos que faltan consiste en sustituirlos por un «valor razonable», aunque esto dependerá del contexto, obviamente, y no siempre será posible.

Por ejemplo, completar valores numéricos: si, por ejemplo, nos falta información sobre el salario de un empleado, podemos adscribirle la media de los valores de los salarios del resto de los empleados de la compañía o de su sección. Ni que decir tiene que este método presenta sus problemas y puede inducir a errores o rebajar la calidad del modelo resultante. Aun así, a veces es todo lo que se puede hacer.

2) Datos redundantes. A veces se repiten tuplas que corresponden al mismo objeto.

Ejemplo de repetición de tuplas

A menudo nos encontramos con situaciones en las que un mismo cliente es dado de alta en varias ocasiones incluso con el mismo número de identificación. Este es un resultado típico de la fusión de datos procedentes de bases de datos diferentes. Una variación son aquellos casos en los que un conjunto de valores correspondientes al mismo objeto recibe identificadores diferentes. Pongamos por caso un cliente que ha sido dado de alta varias veces, pero con identificadores diferentes. Aquí tenemos un ejemplo sencillo que relaciona los clientes y sus datos con los centros de compra donde adquiere sus productos dentro de una cadena de tiendas.

Tabla 3

Identificador	Nombre	Dirección	Centro
24,567	Poch	Roca, 33-1	1
32,456	Martínez	Travesera, 222	2
24,567	Poc	Roca, 33-1	1
33,400	Sala	Diagonal, 556	1
33,441	Arregui	Diagonal, 222	2

No sabemos si el cliente 24,567 se llama «Poc» o «Poch», pero hay muchas probabilidades de que este trate de la misma persona. Claro está, necesitamos herramientas para decidir cuál de las dos interpretaciones es la correcta, o si se trata de un problema de asignación de identificadores. Este es un proceso difícil y muchas veces hay que utilizar herramientas estadísticas solo para poder detectarlo. Es muy normal, sin embargo, en datos que proceden de información voluntariamente (mal) dada por los clientes. Los bancos y otras entidades conocen el porqué de esta conducta de algunos clientes, quienes quieren «despistar» u ocultar los distintos cambios de domicilio.

3) Datos incorrectos o inconsistentes. Caso muy común cuando el tipo de valores que puede recibir un atributo no está controlado porque ha sido declarado como «texto libre», o bien está definido como un tipo determinado (cadena alfanumérica, por ejemplo), pero no se han mantenido los procesos de control de errores necesarios. Por ejemplo, un cliente con una edad superior a cincuenta años que reciba descuentos por Carné Joven; o el cliente que tiene una calle que no corresponde al código postal que tiene asignado; o bien una población que no corresponde al código postal.

Ejemplo de datos inconsistentes

Un ejemplo de datos inconsistentes es el caso siguiente, si suponemos que solo hay diez tiendas en una determinada cadena comercial, es evidente que sucede algo extraño con el cliente «Martín».

Tabla 4

Identificador	Nombre	Dirección	Centro
24,567	Poch	Roca, 33-1	1
32,456	Martínez	Travesera, 222	2
33,345	Martín	Roca, 33-1	144
33,400	Sala	Diagonal, 556	1
33,441	Arregui	Diagonal, 222	2

4) Errores de transcripción. Muy típicos y que pueden dar lugar a alguno de los problemas anteriores. Por ejemplo, mayúsculas/minúsculas, acentos y otros caracteres especiales, etc.

Ejemplo típico de error de transcripción

El método que explore los datos puede decidir que «Barcelona» y «BARCELONA» son dos poblaciones diferentes. Y no digamos ya si se ha introducido «BCN».

5) Datos envejecidos. Ciertos datos se convierten en incorrectos porque no han sido actualizados de la manera adecuada. Este caso puede ser muy complejo de detectar al depender mucho del contexto.

Veamos unos ejemplos de datos envejecidos: un ejemplo típico de esta categoría es el domicilio o la domiciliación bancaria cuando no se notifican los cambios correspondientes. Otro caso podría darse cuando se trata de trabajar con rangos de edades, y en los datos cada persona aparece con el rango de edad correspondiente. Supongamos, por ejemplo, que en lugar de guardar la fecha de nacimiento se guarda la edad del cliente cuando se da de alta. Si no hay un procedimiento de actualización de las edades, lo que ocurre es que la asignación de un cliente a una edad no queda modificada. Por ejemplo, todos los clientes que en el año 2014 tenían cincuenta y nueve años, en el año 2020 tienen o tendrán sesenta y cinco. Han pasado de la categoría de clientes «veteranos» a «jubilados», pero en la base de datos se siguen considerando clientes de la primera categoría.

6) Variaciones en las referencias a los mismos conceptos. Por ejemplo, un abogado puede ser considerado como «profesional liberal», mientras que otro cliente que también lo sea puede estar categorizado como «autónomo». Es más probable que esto suceda si la misma información se guarda en bases de datos diferentes.

Ejemplo de variaciones en las referencias a los mismos conceptos

Aquí tenemos un ejemplo sencillo de la situación que consideramos. Es una simplificación de un caso real del ámbito bancario en el que el banco también ofrece a sus clientes seguros para automóviles. En tal caso, la división de seguros procedía de la adquisición por parte del banco de una compañía de seguros. La rápida fusión de los sistemas de información de ambas empresas provocó, entre otras consecuencias, que durante un largo periodo de tiempo los datos de los clientes comunes se guardaran por duplicado en dos bases de datos diferentes, en las que, además, algunos atributos, como el de Profesión, adoptaban valores de conjuntos diferentes. Aquí lo tenéis. Fijaos en los clientes «Martínez» y «Sala»: sabemos que ambos son profesores universitarios. Sin embargo, para el banco «Martínez» es profesor y para la aseguradora «Sala» era maestro. Como el banco también contemplaba la categoría profesional Maestro entre las que podían ser asignadas a sus clientes, pues pasó lo que pasó...

Tabla 5

Identificador	Nombre	Profesión	Riesgo
24,567	Poch	Abogado	Todo riesgo
32,456	Martínez	Profesor	Terceros
33,345	Martín	Construcción	Todo riesgo
33,400	Sala	Maestro	Todo riesgo
33,441	Arregui	Cocinero	Terceros

Aquí tenemos los datos correspondientes a los tipos de préstamo solicitados por cada cliente y la cantidad correspondiente. ¿Qué le pasa a «Martínez»? ¿Y a «Sala»? ¿Cuál es su verdadera profesión? ¿Cómo lo detectamos?

Tabla 6

Identificador	Nombre	Profesión	Préstamo	Montante	Saldo actual	Saldo medio
24,567	Poch	Abogado	Personal	10,000,000	4,500,000	3,200,000
32,456	Martínez	Profesor	Hipoteca	25,000,000	1,000,000	1,567,000
33,345	Martín	Construcción	Personal	3,000,000	4,000,000	6,563,316
33,400	Sala	Maestro	Hipoteca	6,000,000	2,000,000	5,012,233
33,441	Arregui	Cocinero	Personal	5,500,000	40,000,000	3,245,678

Pues bien, para este tipo de problemas es para el que los *data warehousing* y otros sistemas de gestión de bases de datos intentan aportar soluciones.

7) Datos sesgados. Este tipo de problema puede darse con datos que cumplen todos los otros requisitos de calidad mencionados hasta el momento. Se trata de aquellos tipos de datos que, en conjunto, reflejan preferentemente un valor determinado o conjunto de valores o que proceden de un conjunto de objetos muy determinado. A veces los estudios de minería de datos van precisamente en la dirección de encontrar este tipo de subconjuntos. Otras veces no interesa disponer de este tipo de conjuntos de datos.

Ejemplo de datos sesgados

Es posible que hayamos elegido sin darnos cuenta un conjunto de clientes que son mayoritariamente jóvenes o de determinado tipo de profesión. Según cuál sea el objetivo de nuestro estudio, puede no interesar este tipo de sesgo. Los modelos generalmente intentan explicar o predecir pensando en las mayorías, y pueden ser muy incorrectos para las categorías que no han sido tenidas en cuenta. Además, los sesgos en los datos y los modelos construidos plantean problemas éticos e incluso legales cuando afectan a personas en función de su género, etnia, etc. Existen mecanismos para intentar balancear conjuntos de datos que presentan un sesgo, aunque sería mejor no tenerlo presente de origen.

Suponiendo que después de hacer la «limpieza» hayamos conseguido dejar los datos en un estado de calidad aceptable, todavía hay más cosas por hacer.

2.2.2. Transformación de los datos

No siempre los datos se encuentran en la forma más adecuada para poder aplicar los métodos necesarios para la tarea que hay que llevar a cabo y el modelo que se quiere. En general, nos encontraremos con que tendremos que efectuar alguna de estas transformaciones:

1) Datos numéricos a categóricos. Los datos categóricos son atributos que toman valor en un conjunto finito de etiquetas simbólicas. Por ejemplo, el atributo Edad, para una determinada tarea, puede ser descrito bastante bien como Mayor o Joven porque sean esos los grupos de edad que interesa distinguir y estudiar. Ahora bien, puede suceder que en la base de datos o en varias bases de datos utilizadas este campo posea un valor numérico (edad entre los valores numéricos de dieciocho a cien años, por ejemplo). La solución consiste en asignar una categoría a cada rango de valores que necesitemos, fijando bien

una correspondencia entre los valores numéricos y la categoría (por ejemplo, catagoría Mayor puede corresponder a los valores mayores que sesenta años), o bien intercediendo automáticamente procedimientos de discretización.

2) Datos categóricos a numéricos. Disponemos de datos que aparecen descritos mediante valores categóricos, y lo que necesitamos realmente es disponer de los valores numéricos correspondientes. Tenemos que efectuar el proceso inverso, adscribiendo una traducción a cada categoría en el correspondiente conjunto de valores numéricos. Por ejemplo, haciendo que la categoría Joven del atributo Edad equivalga al rango de valores 18-25 años. El problema es que por cada aparición en la base de datos, quizá no podríamos poner un intervalo, sino un valor único. En este caso hay que efectuar nuevas transformaciones.

3) Otras transformaciones. Muchas veces, para simplificar la representación hay que efectuar otro tipo de transformaciones, ya sea de escala, o de unidades (por ej., en el caso de cantidades relativas a dinero en moneda de diferentes países). Por ejemplo:

a) Simplificación de valores: por ejemplo, dividir los sueldos por mil o un millón.

b) Agrupación de valores continuos en franjas: por ejemplo, todas las compras entre las ocho y las diez corresponden al valor 1, las de diez a doce, al valor 2, etc.

c) Normalización de datos: poner los valores numéricos en un intervalo determinado. Por ejemplo, muchas redes neuronales y algoritmos de agrupación obligan (o prefieren) a que los valores numéricos estén entre 0.0 y 1.0.

d) Adición de una etiqueta que indique a qué clase pertenece un registro. Por ejemplo, en el caso de los seguros, si un cliente pertenece a la clase de riesgo o no (lo cual se puede haber derivado de la experiencia o bien a partir de otro método de minería de datos).

e) Expansión de un atributo: por el hecho de que el valor de un atributo puede adoptar valores en un conjunto limitado de categorías. Por ejemplo, el atributo Riesgo de incendio puede tomar valores en las categorías Alto, Bajo y Medio; y por el hecho de que haya que expresar los datos en forma numérica podemos disgregar el atributo Riesgo de incendio en los atributos Riesgo-alto, Riesgo-medio y Riesgo-bajo, cada uno de los cuales puede tomar el valor 0 o 1 indicando la existencia o no de cada tipo de riesgo. En inglés, esto es lo que se conoce como *dummy variables*.

No todas estas transformaciones –en ausencia de otras herramientas– pueden hacerse utilizando el lenguaje de consulta y manipulación de bases de datos del que se disponga. Cuanto más evolucionadas son las herramientas de minería de datos, más facilidades dan en este sentido y más transparente resulta para el usuario la interacción de la herramienta con el sistema de bases

de datos subyacente. Los *data warehouses* se caracterizan por dar todavía más facilidades en este sentido.

Aquí tenemos un ejemplo en el que podemos ver varias transformaciones de datos. De la tabla original:

Tabla 7

Identificador	Nombre	Profesión	Préstamo	Montante	Saldo actual	Saldo medio
24,567	Poch	Abogado	Personal	10,000,000	4,500,000	3,200,000
32,456	Martínez	Profesor	Hipoteca	25,000,000	1,000,000	1,567,000
33,345	Martín	Construcción	Personal	3,000,000	4,000,000	6,563,316
33,400	Sala	Maestro	Hipoteca	6,000,000	2,000,000	5,012,233
33,441	Arregui	Cocinero	Personal	5,500,000	40,000,000	3,245,678

Expandiendo atributos y aplicando transformaciones numéricas obtenemos la tabla que vemos a continuación:

Tabla 8

Identificador	Nombre	Profesión	Personal	Hipoteca	Montante	Saldo actual	Saldo medio
24,567	Poch	Abogado	1	0	10	4.5	3.2
32,456	Martínez	Profesor	0	1	25	1	1.6
33,345	Martín	Construcción	1	0	3	4	6.6
33,400	Sala	Maestro	0	1	6	2	5.0
33,441	Arregui	Cocinero	1	0	5.5	40	3.2

Otros conjuntos de cambios y transformaciones se originan por motivos diferentes. En efecto, nuestra fuente o fuentes de datos pueden reunir información sobre un cierto conjunto de atributos, y puede que lo que necesitemos sea un conjunto diferente. Comentamos a continuación los problemas y las soluciones más habituales:

1) Derivación de datos: podemos utilizar los atributos de los datos existentes para derivar atributos nuevos (y generar, de hecho, un conjunto nuevo de datos) que nos sean más útiles para el tipo de estudio de minería de datos que se esté llevando a cabo.

Ejemplo de derivación de datos

Típicamente, puede derivarse el atributo Edad de la diferencia existente entre la fecha actual y la fecha de nacimiento declarada. O, para un estudio médico sobre la obesidad, puede ocurrir que dispongamos de los datos siguientes: edad, altura (en m), peso (en kg), sexo y profesión.

Tabla 9

Identificador	Altura	Peso	Sexo	Profesión
24,567	1.90	88	Mujer	Abogado
32,456	1.85	92	Hombre	Maestro
33,345	1.78	73	Hombre	Construcción
33,400	1.70	65	Mujer	Representante
33,441	1.78	110	Hombre	Cocinero

No obstante, nos interesa obtener el índice de masa corporal (IMC), que corresponde a esta sencilla fórmula:

$$IMC = \frac{P}{A^2}$$

donde P es el peso en kilogramos y A es la altura en metros.

Tabla 10

Identificador	Altura	Peso	Sexo	Profesión	IMC
24,567	1.90	88	Mujer	Abogado	24.4
32,456	1.85	92	Hombre	Maestro	26.9
33,345	1.78	73	Hombre	Construcción	23.0
33,400	1.70	65	Mujer	Representante	22.5
33,441	1.78	110	Hombre	Cocinero	34.7

De hecho, al realizar este cálculo para todo el conjunto de datos, lo que conseguimos es reducir el conjunto de atributos original, ya que el IMC es equivalente a la información combinada del peso y la altura. Normalmente, sin embargo, las derivaciones suelen ser más complejas, involucrar más de un atributo y generar también más de un resultado.

Como podemos entender, desde el punto de vista de las bases de datos, la derivación de datos supone crear una relación nueva que es la resultante de incluir un atributo nuevo a la relación original.

2) Fusión de datos o enriquecimiento: puede interesar añadir datos procedentes de otras relaciones o, incluso, de otras bases de datos aportadas desde fuentes distintas.

Ejemplo de enriquecimiento

Podemos añadir a la información que tenemos de nuestros clientes el resultado de una encuesta en la que les hubiéramos preguntado qué coche querrían comprar y si piensan cambiar de coche el próximo año.

Aquí tenemos la tabla de clientes original:

Tabla 11

Identificador	Nombre	Profesión	Préstamo	Montante	Saldo actual	Saldo medio
24,567	Poch	Abogado	Personal	10,000,000	4,500,000	3,200,000
32,456	Martínez	Maestro	Hipoteca	25,000,000	1,000,000	1,567,000
33,345	Martín	Construcción	Personal	3,000,000	4,000,000	6,563,316
33,400	Sala	Representante	Hipoteca	6,000,000	2,000,000	5,012,233
33,441	Arregui	Cocinero	Personal	5,500,000	40,000,000	3,245,678

A continuación, lo que contestaron a una encuesta telefónica:

Tabla 12

Identificador	Tipo de coche	Año próximo
24,567	BMW	Sí
32,456	Skoda	No
33,345	Nissan	Sí
33,400	Mercedes	No
33,441	Smart	Sí

Fusionando las dos tablas en una tabla nueva, operación que en una base de datos relacional puede hacerse con una operación de Join, obtenemos:

Tabla 13

Identificador	Nombre	Profesión	Préstamo	Montante	Saldo actual	Saldo medio	Tipo de coche	Año próximo
24,567	Poch	Abogado	Personal	10,000,000	4,500,000	3,200,000	BMW	Sí
32,456	Martínez	Maestro	Hipoteca	25,000,000	1,000,000	1,567,000	Skoda	No
33,345	Martín	Construcción	Personal	3,000,000	4,000,000	6,563,316	Nissan	Sí
33,400	Sala	Representante	Hipoteca	6,000,000	2,000,000	5,012,233	Mercedes	No
33,441	Arregui	Cocinero	Personal	5,500,000	40,000,000	3,245,678	Smart	Sí

Entonces, podemos responder a preguntas como ¿quién nos pedirá un crédito el próximo año?

2.2.3. Reducción de la dimensionalidad

Una de las justificaciones más frecuentes para el uso de herramientas de minería de datos es su capacidad de trabajar con grandes conjuntos de datos. Ahora bien, el tamaño de un conjunto de datos, o de un problema de minería de datos, lo da tanto la cantidad de registros que tiene como el número de atributos que se manejan. Lo que sucede es que, a partir de ciertos niveles de registros y atributos, la eficiencia de los algoritmos de minería de datos empieza a reducirse. Es la llamada *maldición de la dimensionalidad*. Por lo tanto, si es posible trabajar con menos datos y obtener los mismos resultados, sería mejor desde un punto de vista de eficiencia.

Los métodos de reducción de dimensionalidad buscan justamente trabajar con menos datos y obtener los mismos resultados. A continuación, presentamos brevemente las técnicas habituales de reducción de la dimensionalidad.

1) Reducción del número de registros por tratar. La reducción del número de registros por tratar consiste en encontrar un conjunto de datos de menores dimensiones para construir el tipo de modelo que necesitamos con el nivel de calidad necesario.

La estadística ha desarrollado herramientas para elegir conjuntos suficientes de datos de cara a la construcción de modelos. Así pues, es bueno recurrir a sus técnicas para obtener un conjunto más reducido, pero igualmente potente, de datos iniciales. Los problemas que tenemos que evitar aquí son principalmente que el conjunto elegido no sea demasiado sesgado hacia un conjunto de objetos con características muy concretas y poco representativas, como ya hemos comentado antes. También sucede que, si el conjunto es demasiado reducido, las conclusiones que se extraerán del modelo resultante final no serán lo suficientemente significativas. Por lo tanto, en algunas ocasiones es necesario conservar un conjunto alto de registros para mantener la calidad suficiente.

Una diferencia que hay que destacar de la forma en la que se deben seleccionar las muestras de datos en estadística y en minería de datos es que, mientras que en la primera los casos extremos (*outliers*) se descartan sistemáticamente, en la segunda pueden ser los que más interesan. En consecuencia, los métodos para reducir el número de registros que se utilizarían en el análisis de datos tradicional y en minería de datos son ligeramente diferentes.

Como ejemplo de reducción del número de registros por tratar, supongamos que para predecir algún comportamiento determinado de nuestros clientes quizá no haya que tratar toda la base de datos de clientes, sino una muestra significativa más reducida.

2) Reducción del número de atributos por tratar. La reducción del número de atributos por tratar también recibe el nombre de *selección de atributos*.

Este hecho representa tener que detectar atributos irrelevantes que carecen de efecto alguno sobre la calidad del modelo final (es decir, que el modelo nos permite contestar las mismas preguntas con o sin esos atributos), y atributos o combinaciones de atributos equivalentes (atributos que permiten hacer el papel de grupos de otros atributos sin afectar a la calidad final del modelo).

Los métodos de selección de atributos tienen una cierta complejidad y los introduciremos de la manera adecuada con un par de ejemplos.

Ejemplo

Veamos un ejemplo sencillo. Supongamos que tenemos los datos de obesidad de una serie de clientes del banco:

Tabla 14

Identificador	Altura	Peso	Sexo	Profesión	IMC
24,567	1.90	88	Mujer	Abogado	24.4
32,456	1.85	92	Hombre	Maestro	26.9
33,345	1.78	73	Hombre	Construcción	23.0
33,400	1.70	65	Mujer	Representante	22.5
33,441	1.78	110	Hombre	Cocinero	34.7

Como hemos visto, parece razonable pensar que la combinación de atributos (Altura, Peso) es equivalente al IMC, ya que este último atributo se calcula a partir de los otros dos. Podemos decir por lo tanto que la combinación de atributos (Altura, Peso) aporta la misma información que el atributo IMC, con lo cual podemos sustituir los dos atributos Altura y Peso por el atributo IMC con una sencilla operación SELECT típica de las bases de datos relacionales.

Tabla 15

Identificador	Sexo	Profesión	IMC
24,567	Mujer	Abogado	24.4
32,456	Hombre	Maestro	26.9
33,345	Hombre	Construcción	23.0
33,400	Mujer	Representante	22.5
33,441	Hombre	Cocinero	34.7

Con un ejemplo tan sencillo como este es evidente que no ahorramos mucho, pero en

bases de datos de millones de clientes es importante realizar un análisis previo a fin de encontrar este tipo de equivalencias. Cada atributo menos puede representar millones de bytes necesarios para representar el conjunto de datos.

Ejemplo

Ahora veremos otro ejemplo no tan directo y que tiene que ver con los atributos irrelevantes. Volvemos a nuestra base de datos del banco imaginario. Hemos añadido un atributo nuevo que corresponde a la clase Cliente. Cada cliente no puede pertenecer sino a una sola clase. La clase 0 es la de los clientes de poca morosidad; la clase 1, la de alto riesgo de morosidad. No es necesario que nos preocupemos ahora de cómo se ha obtenido esta clasificación. Además, hemos introducido algunos cambios en los valores para dejar más claro lo que queremos decir:

Tabla 16

Ident.	Nombre	Profesión	Préstamo	Montante	Saldo actual	Saldo medio	Tipo de coche	Año próximo	Clase
24,567	Poch	Abogado	Personal	10	4.5	3.2	BMW	Sí	0
32,456	Martínez	Maestro	Hipoteca	25	1	1.6	Skoda	No	1
33,345	Martín	Construcción	Personal	3	4	6.6	Nissan	Sí	0
33,400	Sala	Representante	Hipoteca	6	2	5.0	Mercedes	No	1
33,441	Arregui	Cocinero	Personal	5.5	40	3.2	Smart	Sí	0

Aquí hay dos atributos que, aun pareciendo interesantes, nos presentan un problema. Todo el mundo que tiene un préstamo personal ha contestado afirmativamente a la pregunta de si se comprará un coche el año próximo; todo el mundo que tiene una hipoteca ha contestado negativamente. Por lo tanto, estos dos atributos son muy poco informativos. Evidentemente, si hubiera más tipos de préstamos, esto no sería así, aunque entonces lo que sucedería es que tendríamos unos datos insuficientes. Está claro que eso tendríamos que matizarlo. Si queremos establecer asociaciones o dependencias, parece que hay una fuerte dependencia entre el tipo de préstamo y la intención de compra, lo cual ya es bastante significativo. Si queremos agrupar a los clientes por su similitud, quizá estos dos atributos no nos hacen falta para nada.

2.3. Minería de datos: el proceso de construcción de modelos

En la fase de minería de datos tenemos los datos con la calidad adecuada, en el formato adecuado y hemos seleccionado los atributos y los registros aparentemente necesarios y relevantes. Tenemos decidido qué tipo de modelo queremos obtener. Por lo tanto, ahora hay que elegir un método de construcción de modelos entre la multitud de métodos que permiten obtener el tipo de modelo que nos interesa.

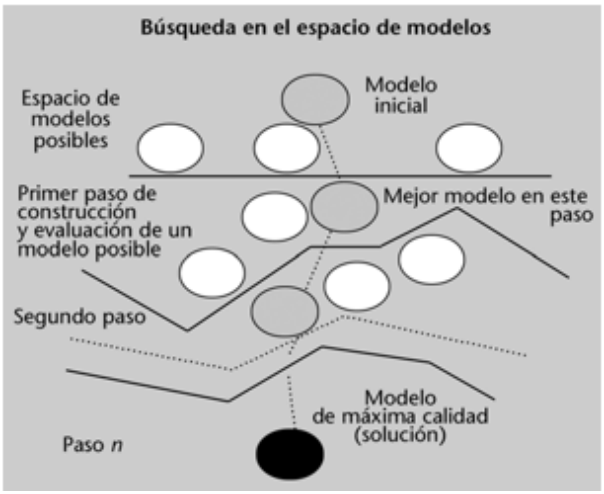
De hecho, para una misma tarea varios modelos pueden sernos útiles. Igual que la aparente diversidad de modelos puede ocultarnos las similitudes con la tarea para la que tiene que aplicarse, tampoco podemos dejar de remarcar la gran similitud existente en el proceso de construcción de los distintos tipos de modelos.

En efecto, el proceso de construcción de modelos consiste en encontrar el modelo (el conocimiento) que responde mejor a las características implícitas en los datos. Este tipo de problema suele conceptualizarse como un proceso de búsqueda.

Un proceso de búsqueda consiste en explorar un espacio de modelos posibles (por ejemplo, en encontrar la mejor red neuronal entre todas las posibles contando para ello con una serie de datos de entrada y salida) para encontrar el que tenga la mejor calidad.

Podemos representar el proceso de búsqueda con el esquema de la figura 7.

Figura 7. Búsqueda en el espacio de modelos



2.3.1. Mecánica general del proceso de búsqueda

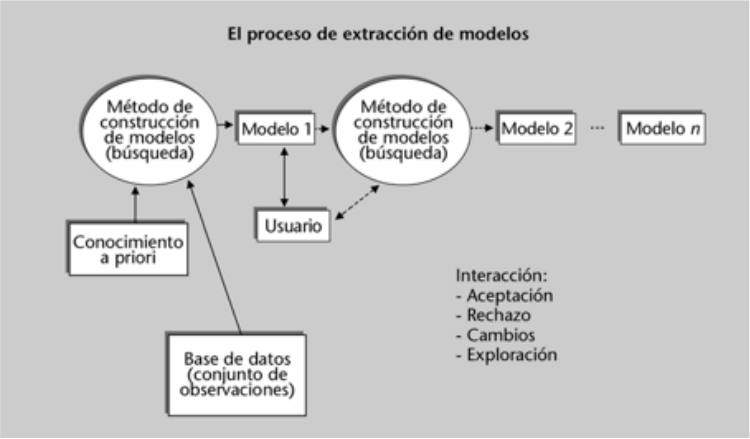
¿En qué consiste un proceso de búsqueda, entonces? Todo proceso de búsqueda parte de un modelo inicial (que puede ser el modelo vacío, inexistente) y en cada paso modifica este modelo mediante un conjunto de operadores de modificación de modelos. En cada paso es posible aplicar más de un operador, y también se pueden generar varios modelos nuevos alternativos:

- Si alguno de estos modelos ya tiene la suficiente calidad, entonces se puede considerar el modelo final y hemos encontrado la solución.

- Si el modelo no es todavía la solución, hemos de elegir alguno de entre los demás y continuar aplicando operadores de transformación hasta que encontremos una solución o hasta que no haya más combinaciones posibles por obtener.

Puesto que la mayor parte de las veces el conjunto de combinaciones posibles es tan grande que resulta imposible generarlas todas en un tiempo razonable, hay que elegir acertadamente en cada paso un subconjunto de modelos parciales. Así, seleccionaremos modelos que parezca que aseguran un modelo final. Para saber cuál es el mejor modelo, utilizaremos algún tipo de función de evaluación, que da a un modelo parcial un valor más alto cuanto más probable sea que se encuentre en el camino hacia un modelo final bueno, de calidad alta.

Figura 8. El proceso de extracción de modelos



No hay que ocultar que los principales problemas de este tipo de conceptualización son los siguientes:

1) Disponer de una medida de calidad que califica como el mejor posible un modelo que solo es relativamente bueno con respecto a los modelos que tiene cerca de él, pero que no es el modelo óptimo dentro de todo el espacio de soluciones posibles. Este problema se denomina *problema de la obtención de mínimos locales*. Sería como elegir la manzana más grande de un frutero, sin tener en cuenta que puede haber manzanas aún más grandes en otros fruteros.

2) Obtener un modelo que sea bueno teniendo en cuenta solo los datos de los que disponemos. Por poner un ejemplo, si estamos modelizando el comportamiento de los clientes de una empresa y disponemos de una base de datos de tres mil clientes para hacer pruebas, de la que obtenemos un modelo clasificatorio, no queremos que al ver el caso tres mil uno el modelo se equivoque y lo asigne a la clase que no le corresponde (por ejemplo, no queremos que un cliente con riesgo de morosidad sea clasificado como no moroso). Este problema que consiste en obtener modelos que generalizan mal y son demasiado específicos con respecto a los datos que han visto se conoce como el *problema de la sobreespecialización* (en inglés, *overfitting*).

3) En el transcurso del tiempo es posible que el modelo obtenido se degenere porque empiezan a aparecer observaciones nuevas para las que no generaliza bien. Este es el problema del envejecimiento de modelos. Es necesario actualizarlos, ya sea desde cero o incrementalmente, dependiendo del tipo de modelo del que se trate.

Vemos entonces que todos los procesos de construcción de modelos se distinguen por las características siguientes:

- El lenguaje de descripción. Es decir, aquello que expresan los modelos (asociaciones, dependencias, similitudes, etc.).

- Los operadores de modificación de modelos parciales. Es decir, cómo se van construyendo los modelos.

- La función de evaluación. Evalúan que el modelo que se está construyendo es mejor o peor, permitiendo comparar modelos.

2.3.2. Variedad de modelos de búsqueda

Por lo que acabamos de decir, es como si todos los métodos de construcción de modelos funcionaran de la misma manera: parten de los datos y avanzan por el espacio de soluciones guiándose por su función de evaluación. Bien, esto es verdad solo hasta cierto punto y es recomendable en parte. De hecho, si recordamos que el modelo tiene que ser comprensible y utilizable por alguien que toma decisiones, conviene ver si el proceso admite también la presencia de un observador humano que en todo momento del proceso de búsqueda pueda evaluar cómo ha ido funcionando el proceso y hacer que la búsqueda adopte una cierta dirección.

En el fondo, puede considerarse que un proceso de búsqueda parte de más información que los datos. Veámoslo esquematizado en la figura 8. Los elementos por considerar en este esquema son los siguientes:

1) Base de datos. Conjunto de observaciones o ejemplos de los que disponemos.

2) Conocimiento *a priori*. Todo aquel tipo de conocimiento del que se dispone y se quiere utilizar. Por ejemplo, saber que determinadas pautas de conducta siempre están asociadas. En el caso del banco del que hemos estado hablando, podemos expresar el conocimiento de que los clientes de renta alta viven mayoritariamente en un barrio determinado y, por lo tanto, indicamos que tenemos preferencia por modelos que reflejen ese hecho. Aquí hay un aspecto importante: si los datos contradicen este conocimiento *a priori*, ¿qué debemos tener en cuenta? ¿Los datos o el usuario humano que tiene que dar esta indicación inicial? ¿Mitad y mitad? ¿Cómo alteramos el modelo que se está construyendo con el fin de equilibrar la influencia de las dos fuentes de información, el usuario y los datos?

¿Cómo podemos expresar este conocimiento *a priori*? Aquí exponemos algunas de las técnicas de expresión de este conocimiento más típicas utilizadas en algunos de los métodos, aunque no siempre aplicables a todos ellos:

- Distribución *a priori* de la probabilidad de los valores de los datos: normalmente, es algún parámetro que permite fijar las características de la distribución que siguen los datos. Por ejemplo, podemos suponer que la

distribución que siguen los datos es una normal multivariante y determinar los parámetros que creemos importantes.

- Relaciones entre los datos: indicar que consideramos que dos atributos han de ser tratados como dependientes, por ejemplo. En el ejemplo que hemos comentado más arriba, los atributos serían el nivel de renta y el barrio de residencia.

- Relaciones entre los datos y parte del modelo: por ejemplo, en el caso de agregación, forzar que determinadas observaciones se encuentren en el mismo grupo de observaciones final.

- Relaciones de coocurrencia: parecidas a las de relaciones de dependencia que hemos comentado antes. Por ejemplo, «siempre que alguien compra cerveza, también compra cacahuetes».

3) Control del proceso de búsqueda. En principio, el proceso de búsqueda se guía por las características de calidad de los modelos que va construyendo, o bien por una medida de calidad que es una combinación del grado con el que el modelo se ajusta a los datos y las preferencias del usuario. Esto no quiere decir que el proceso de búsqueda proceda automáticamente hasta encontrar un modelo de la suficiente calidad. El usuario puede decidir detener el proceso de búsqueda y retroceder hasta una situación anterior en la que empieza a explorar introduciendo otra información. Este era el sistema seguido, por ejemplo, por Data Surveyor.

Lectura recomendada

Encontraréis información sobre la propuesta original de Data Surveyor en la obra siguiente:
M. Holsheimer; A. Siebes (1994, enero). *Data mining: the Search for Knowledge in Databases*. Reporte técnico CS-R9406. Ámsterdam: Centrum voor Winskunde en Informatica (CWI).

Según este esquema, pues, podemos dividir los métodos de minería de datos en lo que concierne a conocimiento *a priori*, al tipo de datos y al proceso de construcción, en los términos que anotamos a continuación:

1) En cuanto a conocimiento *a priori*:

- Tipo de conocimiento *a priori* que permiten expresar.

- Lenguaje en el que permiten expresar ese conocimiento.

- Reutilización del conocimiento extraído por otro sistema de minería de datos con el fin de guiar el proceso de búsqueda. En inglés, *multistrategy learning*.

2) Por lo que respecta al tipo de datos:

- Métodos que utilizan solo observaciones, es decir, solo información con respecto a los valores que se han recogido, sin más información añadida. Los métodos que parten de esta base se conocen como *métodos no supervi-*

sados. Un ejemplo típico de ello son los métodos de agregación. En inglés, *clustering*.

- Métodos que añaden información discriminando las observaciones. Por ejemplo, a cada observación se le asigna el nombre o el indicativo de la clase a la que pertenece. Hablamos propiamente, entonces, de ejemplos y contraejemplos, más que de observaciones. Los métodos que parten de esta situación inicial se denominan *métodos supervisados*. Un ejemplo típico de ello es la clasificación.

3) En lo que concierne al proceso de construcción:

- Métodos *batch*. Métodos que están pensados para utilizar todos los datos existentes como único conjunto de datos para obtener un modelo «en un solo paso».

- Métodos incrementales. Aquellos métodos que están pensados para ir construyendo un modelo con porciones del conjunto total de observaciones (incluso de observación en observación) sin guardar memoria de todas las observaciones vistas anteriormente. Útiles para ir modificando los modelos a medida que van cambiando los datos en el tiempo (envejecimiento de datos o deriva de modelos).

- Métodos interactivos. Métodos en los que el usuario posee un papel en cada paso o serie de pasos que efectúa el procedimiento, introduciendo conocimiento nuevo y dando indicaciones sobre hacia dónde hay que dirigir la búsqueda antes de construir el modelo final.

Lectura recomendada

El trabajo siguiente describe aspectos metodológicos sobre el proceso de minería de datos:
A. Feelders; H. Daniels; M. Holsheimer (2000). «Methodological and practical aspects of data mining». *Information and Management* (vol. 37, núm. 5, págs. 271-281).

2.3.3. Evaluación e interpretación del modelo

El final de la fase de minería de datos es un modelo que representa un tipo determinado de conocimiento sobre el dominio que estábamos estudiando. Pero ¿qué calidad presenta este modelo? ¿Podría ser mejor? ¿Podemos considerarlo lo suficientemente bueno, o tenemos que volver a empezar? No hay medidas absolutas y generales de calidad para todo tipo de modelos, ya que, como dijimos, cada modelo se dirige a un objetivo diferente.

Típicamente, el proceso de evaluación consiste en disponer de dos conjuntos de datos procedentes del mismo conjunto inicial (y ya preparado): un conjunto de datos que se utiliza para construir el modelo y otro, para evaluarlo. En algunas ocasiones, se introduce un tercer conjunto entre estos dos: el conjunto de validación. Utilizamos un conjunto de datos para construir el modelo; otro, para darlo por bueno (validarlo), y un tercero, para evaluarlo. Normalmente, estos tres conjuntos, aun reflejando informaciones sobre los mismos atributos, proceden de conjuntos de datos diferentes.

Con todo, hemos de tener en cuenta que no se pueden comparar modelos predictivos con modelos descriptivos. Para los primeros, tenemos que usar medidas que nos indiquen hasta qué punto son buenos haciendo predicciones. Para los segundos, hemos de medir hasta qué punto se ajustan al dominio descrito.

Pero ¿qué medidas se utilizan más a menudo para evaluar la calidad del modelo? Y ¿cómo puede efectuarse la evaluación de una manera más metódica que la que acabamos de apuntar?

Veamos un ejemplo de proceso de evaluación: presentaremos un caso del mundo de los seguros. Si queremos extraer un modelo predictivo que nos indique a partir de una serie de datos de los clientes si un cliente nuevo puede ser de riesgo (tener un número excesivo de accidentes que la aseguradora deba pagar), separaremos la base de datos de clientes –de quienes sabemos cuáles son de riesgo y cuáles, no– en dos bases de datos: una para construir el modelo y otra para validarlo.

Sobre la primera serie de datos aplicamos un método de predicción (por ejemplo, una clasificación que nos indique si un cliente pertenece a la clase de riesgo o a la de no riesgo), y a partir de ahí extraemos aquellas combinaciones de valores que predicen la pertenencia a una clase o a otra.

Por ejemplo, supongamos que los clientes con edad baja (jóvenes) que tienen coches rojos y un carné de conducir con más de dos años de antigüedad son los más propensos a pertenecer a la clase de riesgo. Podemos utilizar esta regla de predicción sobre el conjunto de datos de evaluación para saber si este «minimodelo» es correcto:

- Si con la citada regla se clasifica correctamente un porcentaje de clientes significativo (pongamos el 95 %), entonces podemos considerar que contamos con un modelo de buena calidad. Es decir, si el modelo clasifica correctamente a clientes que el método no había utilizado para construirlo. Pues bien, esto querrá decir: si indica como propensos a ser de riesgo clientes que tenemos etiquetados como de riesgo, y como de no riesgo, clientes que tenemos etiquetados como tales.

- Si, en cambio, la proporción de «falsos positivos» (clientes de no riesgo que se clasifican como de riesgo) o «falsos negativos» (clientes clasificados como no riesgo, cuando en realidad sí son de riesgo) es alta, entonces tenemos un modelo predictivo de baja calidad.

En este último caso tenemos que pensar que algo ha ido mal en todo el proceso de minería de datos (la calidad de los datos, una muestra insuficiente o sesgada, etc.) y reconsiderar los pasos correspondientes. Este pequeño ejemplo nos ha servido para mencionar una posible medida de calidad: el error en la predicción.

Aún nos queda la interpretación final. Cuando ya tenemos un modelo que posee el nivel de calidad requerido y que ha sido validado mediante el proceso de evaluación, hay que interpretarlo y extraer el significado del conocimiento que nos está mostrando. Es aquí donde corremos el riesgo de las falsas interpretaciones.

Un ejemplo muy sencillo de falsas interpretaciones es el de aquel sistema de predicción que determinó que el factor más importante para calcular si una persona residente en Londres podía quedarse embarazada era el sexo, ya que en el 99.99 % de los casos las personas que se quedaban embarazadas eran mujeres, mientras que el 0.01 % restante no se identificaba como tales, indicando que una variable binaria quizás no es adecuada.

Son numerosos los ejemplos de modelos que dan conclusiones evidentes, y hay que estar al acecho en esta fase de interpretación para no recoger como un gran hallazgo algo que ya se sabe. De aquí la importancia de los métodos que permiten integrar conocimiento *a priori*, especialmente conocimiento negativo, de lo que se está seguro que no puede pasar o que no es relevante. No obstante, hay problemas que no son evidentes.

Veamos otro ejemplo, conocido como el caso de la apendicectomía beneficiosa: ilustramos el problema de las falsas interpretaciones con el caso discutido por Wen sobre la interacción entre determinados tipos de operaciones quirúrgicas y la tasa de mortalidad en un hospital público de Ontario (Canadá) entre 1981 y 1990.

Wen se concentró en los casos de pacientes sometidos a una colecistectomía primaria abierta. Algunos de estos pacientes también habían sido sometidos a una apendicectomía en el proceso de colecistectomía, lo que se conoce como una *apendicectomía incidental o discrecional*.

En la tabla siguiente podemos ver los resultados que reflejan las muertes que tuvieron lugar en el hospital comparando los pacientes que habían sufrido apendicectomía durante la operación de colecistectomía primaria abierta y los que no:

Tabla 17. Caso de la apendicectomía beneficiosa

	Con apendicectomía	Sin apendicectomía
Porcentaje de muertes ocurridas en el hospital	21 (0.27 %)	1,394 (0.73 %)
Porcentaje de pacientes supervivientes en el hospital	7,825 (99.73 %)	190,205 (99.27 %)

Se efectuó un test de significación para comparar los resultados de los dos grupos y averiguar si mostraban una diferencia significativa. Se encontraron con que, según el test, la diferencia era, efectivamente, significativa.

Este «descubrimiento» del conocimiento del hecho de que una apendicectomía incidental durante la operación de colecistectomía puede «mejorar» las

probabilidades de sobrevivir hay que tomárselo con un poco de calma. ¿Cómo es posible que una apendicectomía incidental pueda mejorar los resultados?

Wen consideró por separado un grupo de pacientes de bajo riesgo. Este grupo de pacientes, en cambio, mostraba que el efecto de la apendicectomía discrecional presentaba resultados bastante insatisfactorios. Paradójicamente, podría ser que la apendicectomía casual afectase negativamente tanto a los pacientes de bajo riesgo como a los de alto riesgo, pero que, considerados juntos, diera la impresión de un efecto positivo. Esto se conoce como *paradoja de Simpson*, y hay que estar muy al acecho. Veamos cómo acaba de funcionar.

En la tabla siguiente se muestran datos ficticios que permitirían interpretar esa paradoja:

Tabla 18. División en pacientes de bajo y alto riesgo

	Con apendicectomía		**Sin apendicectomía**	
	Bajo riesgo	Alto riesgo	Bajo riesgo	Alto riesgo
Muertes	7	14	100	1,294
Supervivientes	7,700	125	164,009	26,196

En la tabla siguiente hallaremos las proporciones correspondientes a las muertes dentro del hospital clasificadas como apendicectomía incidental y pacientes de riesgo correspondientes con los datos de la tabla anterior:

Tabla 19. Efecto combinado

	Con apendicectomía	**Sin apendicectomía**
Bajo riesgo	0.0009	0.0006
Alto riesgo	0.1000	0.0500
Combinado	0.003	0.0070

Podemos decir que las categorías de riesgo y las muertes se encuentran altamente correlacionadas. Era más probable que las apendicectomías fueran aplicadas a pacientes de bajo riesgo que a los de alto riesgo. Por lo tanto, si no se conoce la categoría de riesgo (relacionada con la edad) de un paciente, pero se sabe que ha pasado por una apendicectomía, entonces podemos decir que es más probable que pertenezca a la categoría de «Bajo riesgo» (jóvenes). Ahora bien, este hecho no implica de ninguna de las maneras que pasar por una apendicectomía disminuya el riesgo de algunos pacientes. Si la información sobre el riesgo no aparece en la tabla, se puede extraer esta conclusión puramente ilusoria.

Wen realizó un estudio de regresión teniendo en cuenta más variables (edad, sexo, situación de entrada en el hospital), y concluyó que no existe forma alguna de afirmar que la mejora a corto plazo se pueda considerar debida a la apendicectomía.

Así pues, la interpretación requiere mucha precaución y más de una mirada sobre los resultados. Por eso conceptos como el mencionado AutoML están lejos de ser realmente una herramienta para construir modelos de minería

Lectura recomendada

Encontraréis el caso estudiado por Wen de la apendicectomía beneficiosa en la obra siguiente:
S. W. Wen; R. Hernández; C. D. Naylor (1995). «Pitfalls in Nonrandomized Studies: The case of incidental Appendectomy with Open Cholecystectomy». *Journal of the American Medical Association* (núm. 275, págs. 1687-1691).

de datos de manera automática, siempre se requerirá la validación por los expertos del ámbito de conocimiento.

Lectura recomendada

Un trabajo reciente que describe los errores típicos que se pueden cometer en la interpretación de modelos es el siguiente:
C. Molnar; G. König; J. Herbinger; T. Freiesleben; S. Dandl; C. A. Scholbeck y otros (2020). «Pitfalls to avoid when interpreting machine learning models». arXiv preprint arXiv:2007.04131

2.4. Integración de los resultados en el proceso

El último paso consiste en integrar los resultados de la minería de datos en el proceso típico del sistema de información en el que esté aplicándose.

Un ejemplo sencillo es el del proceso de documentación textual utilizado en algunos grandes diarios. Cada día, las noticias se clasifican por varias categorías, de manera que los usuarios de los servicios de información de estos diarios pueden hacer consultas por distintas palabras clave. Está claro que detrás de todo esto hay un esfuerzo previo de clasificación. Aplicando algoritmos de minería de datos textuales es posible construir un procedimiento de clasificación que asigne automáticamente las palabras clave a las distintas noticias del diario. La integración correspondería aquí a la transformación del modelo de clasificación en un programa más de la cadena: edición, etiquetado e inclusión en la base de datos documental del diario.

Lectura recomendada

Encontraréis más información sobre los procesos de documentación textual que eran utilizados en algunos grandes diarios en la obra siguiente:
J. Schmitz; G. I. Armstrong; J. D. C. Little (1990). «CoverStory-Automated News Finding in Marketing». *DSS Transactions*.
Actualmente, el ámbito del *text mining* ha evolucionado muchísimo, como muestra el trabajo siguiente:
C. C. Aggarwal; C. Zhai (2012). «A survey of text classification algorithms». En: *Mining text data* (págs. 163-222). Boston, MA: Springer.

No todos los modelos pueden integrarse con facilidad. La mayoría de ellos requieren una transformación al código de programación correspondiente. Buena parte de los sistemas comerciales de minería de datos ofrecen la posibilidad de traducir el modelo obtenido en procedimientos al lenguaje de programación correspondiente e insertarlo después dentro de un tratamiento de información más general.

2.5. Observaciones finales

Pondremos énfasis en que, a pesar de esta presentación lineal, el descubrimiento avanza de manera iterativa. No acaba con la construcción de un modelo y con la generación de información resumida. Una vez que se dispone de un modelo, hay que trabajar con él, hacer preguntas nuevas, preguntarse qué ocurre si en lugar de disponer de las relaciones que muestra el modelo se dieran otras, etc. Este hecho puede representar, a la vez, la aportación de datos que no necesitábamos inicialmente y, por tanto, tener que emprender un proceso de selección y limpieza de datos nuevos que darán, con los datos actualmente existentes, un modelo nuevo, etc.

Alcanzar esta posibilidad de mantener un procedimiento abierto de descubrimiento no es trivial; hay que prever los mecanismos que permitan redefinir fácilmente los datos de interés, transformarlos, etc. Es necesario, pues, contar con una disposición activa para anticipar los requisitos de datos de cada área de interés posible, la conexión de las fuentes de datos adecuados.

3. Las herramientas de minería de datos y las áreas relacionadas

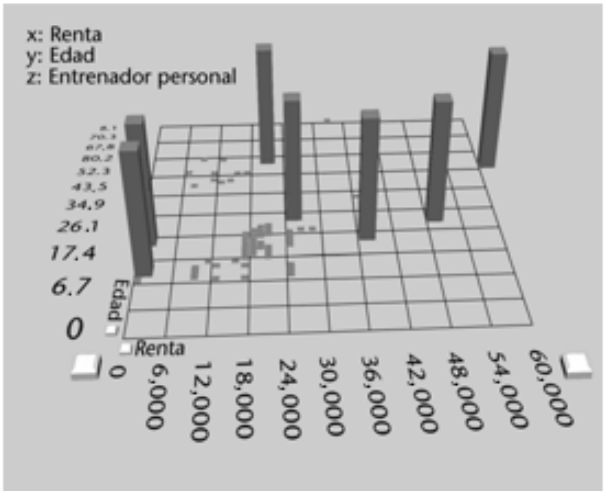
El núcleo de este material docente se centra en los modelos de minería de datos, los métodos para construirlos y los algoritmos sobre los que se basan. No obstante, normalmente hay una serie de herramientas que también guardan relación con la minería de datos y que por lo menos debemos tener presente.

3.1. Herramientas de visualización

Una manera muy potente e intuitiva de obtener conocimiento a partir de datos es mediante la inspección visual, aprovechando las capacidades del sistema visual humano.

Veamos un ejemplo sencillo de la potencia de las herramientas de inspección visual: en la figura 9 aparece representada la relación entre el nivel de ingresos (Renta, eje X) de los socios del club deportivo que utilizaremos como ejemplo a lo largo del programa, su edad (Edad, eje Y) y si solicitan un entrenador personal (Entrenador personal, eje Z).

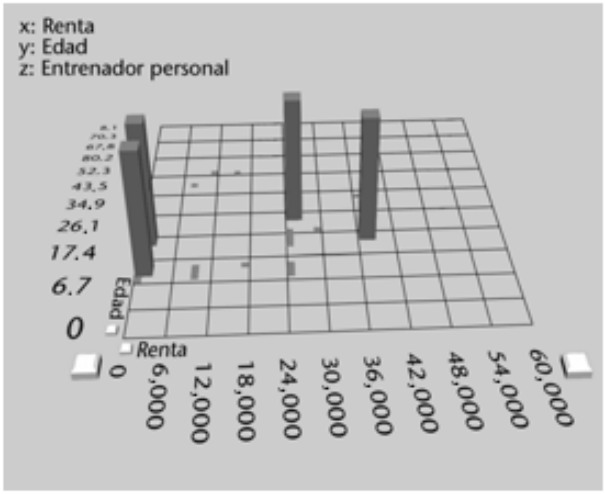
Figura 9



Podemos extraer algún conocimiento de esta visualización: parece que la mayor parte de los socios que solicitan el servicio de entrenador personal se concentran en las rentas superiores a 24,000 euros, incluso una visualización tan poco elegante como esta ya nos permite identificar patrones y tendencias. Hacemos unas cuantas comprobaciones más:

1) Hacemos un filtrado de los datos por sexo, de manera que en el gráfico aparezcan solo los hombres:

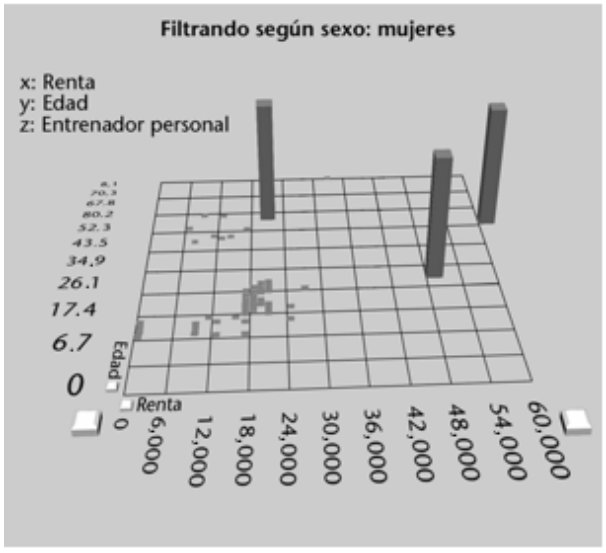
Figura 10



Podemos extraer la conclusión de que quienes solicitan predominantemente el servicio de entrenador son hombres jóvenes con rentas medias-bajas.

2) ¿Y las mujeres? Hacemos un filtrado de los datos para ver solo a las mujeres:

Figura 11. Filtrando según sexo: mujeres

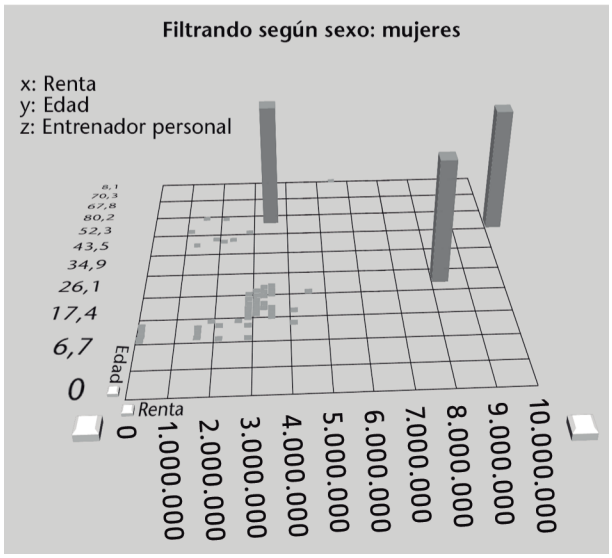


Parece, pues, que se trata de mujeres mayores y con rentas más altas, ¿verdad? Si ahora filtramos lo que tenemos según el distrito de residencia y solo se eliminan los socios que proceden de la zona A (un barrio de clase alta) resulta que casi no aparecen socios en la figura 12.

Así pues, parece que hasta ahora llegaremos a conocer bastante bien a los clientes que solicitan el servicio de entrenador personal en este gimnasio: son hombres jóvenes de renta baja-media, pero que viven en el distrito alto de la ciudad, o bien mujeres mayores con renta media-alta que también viven en el mismo barrio. Ahora, cuando ya tenemos una primera idea de los datos, podríamos aplicar otros métodos que nos dieran un resultado numérico más

preciso y un modelo de predicción más detallado que el que nos permite la inspección visual.

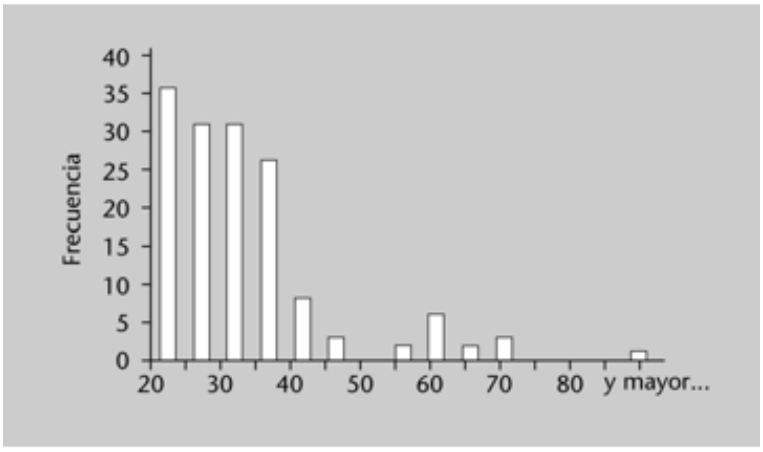
Figura 12



Las herramientas de visualización son bastante útiles en la fase de preparación de datos y la de interpretación de los modelos resultantes. En la primera fase, estas herramientas permiten ayudarnos a conocer mejor los datos, y se complementan con las herramientas típicas de estadística descriptiva, que nos permiten encontrar los valores más frecuentes, la dispersión de valores de cada variable, valores mínimos y máximos, valores muy poco frecuentes y algún tipo de correlación entre las variables consideradas.

Una de las herramientas visuales más tradicional en esta parte son los histogramas. Veamos un ejemplo de utilidad de los histogramas. Aquí podemos ver la distribución de los valores de las edades entre los clientes de nuestro club:

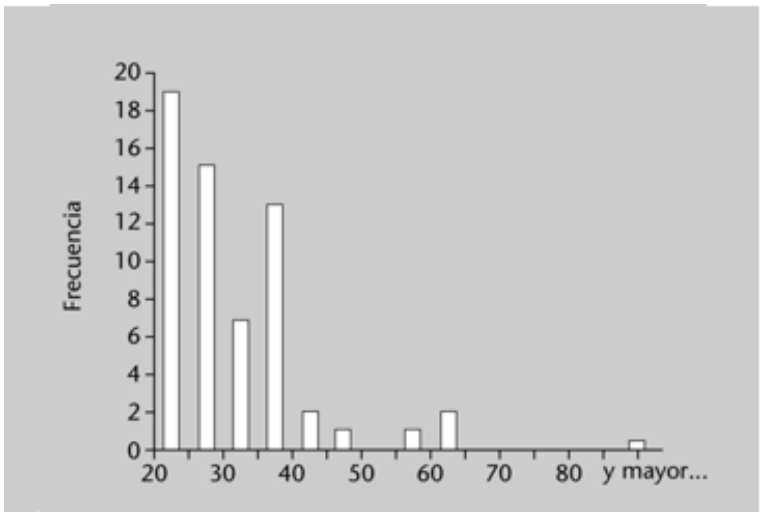
Figura 13



Eligiendo una variable discriminante (o una variable de clase) pueden darse comparaciones entre las distribuciones de valores para varias clases o combi-

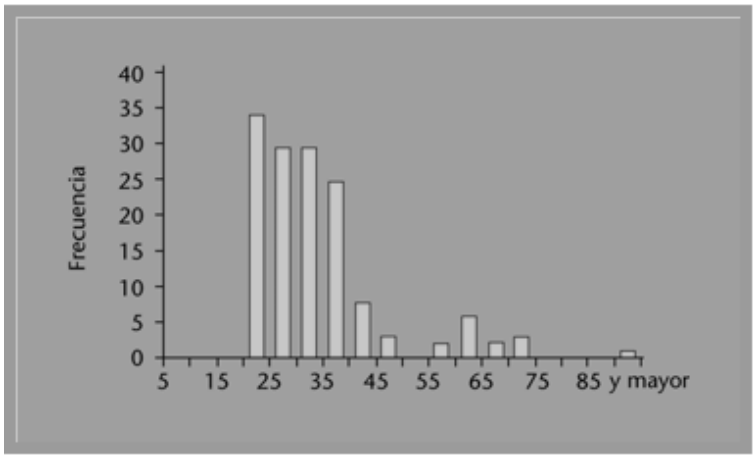
naciones de los valores de los otros atributos. Aquí tenemos la distribución de las edades entre las mujeres que acuden al club:

Figura 14



Y aquí, entre los hombres:

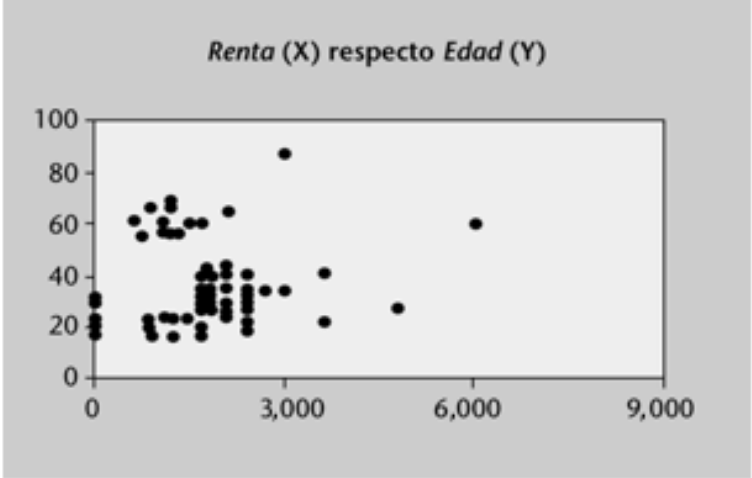
Figura 15



A partir de estas gráficas, ¿podemos decir que hay diferencias significativas? Aunque ambas gráficas no pueden superponerse y compararse directamente, porque, por ejemplo, no comparten la misma escala, sí que permiten observar algunos hechos que pueden ser relevantes (por ejemplo, para los usuarios de más edad).

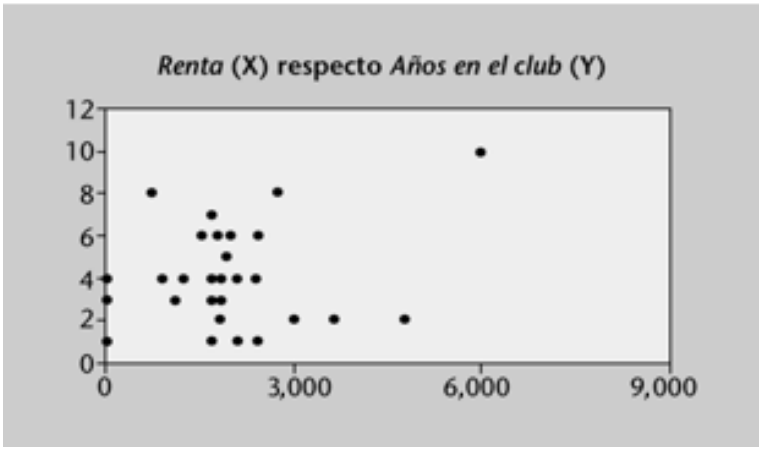
Otra herramienta útil son los diagramas de dispersión, en inglés, *scatterplots*, que dan una idea de la relación existente entre los valores de dos variables. Veamos un ejemplo de utilidad de los diagramas de dispersión. Aquí podemos ver la relación con los valores de las variables Renta y Edad:

Figura 16



Y aquí, entre Renta y Años en el club:

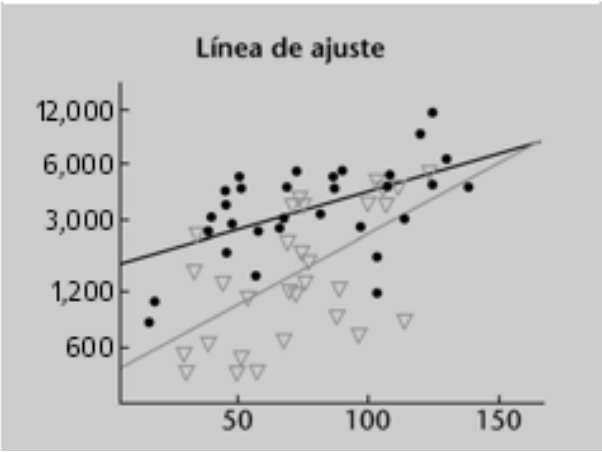
Figura 17



Normalmente, la existencia de relación entre los valores de dos variables podemos estudiarla construyendo una función que dibuje una línea que explique los valores de una variable en función de los de la otra. Asimismo, se puede derivar el coeficiente de correlación entre ambas variables, o de recta de ajuste o de regresión. Ahora no entraremos en definir ni explicar estos conceptos; solo nos centramos en el aspecto de visualización.

Veamos un ejemplo de utilidad de una función de ajuste: aquí presentamos una gráfica de ejemplo en la que se ha encontrado una línea de ajuste que indica una relación funcional entre las variables Ingresos y Tiempo trabajado en tres meses para dos grupos diferentes.

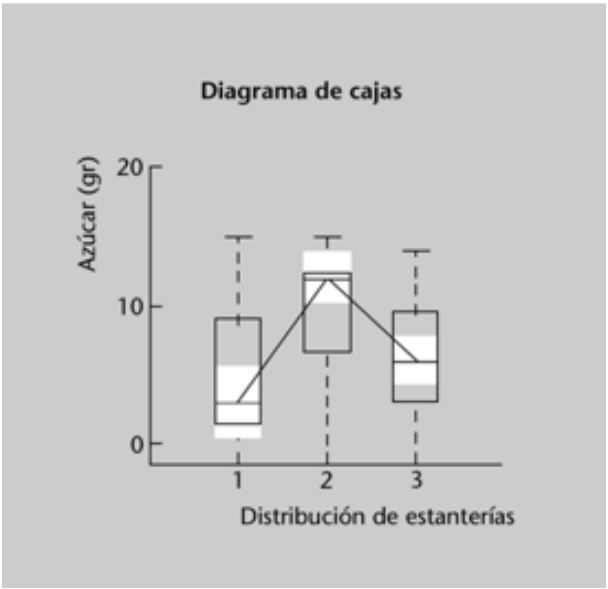
Figura 18. Línea de ajuste



Otra herramienta gráfica que informa de la concentración de valores en torno a un punto son los diagramas de cajas o *boxplot*.

Veamos un ejemplo de utilidad de un diagrama de cajas. Aquí tenemos una muestra de un diagrama de cajas que relaciona determinados productos con su disposición en las estanterías del supermercado:

Figura 19. Diagrama de cajas



El problema de los datos con dimensionalidad elevada (con un conjunto de atributos grande) es que no podemos visualizar completamente las relaciones con todas las variables de manera simultánea. Así, hay que proyectar conjuntos de muchas variables sobre representaciones gráficas de dos o tres dimensiones y agotar las distintas combinaciones de variables dos a dos para preguntarnos sobre los fenómenos de interés.

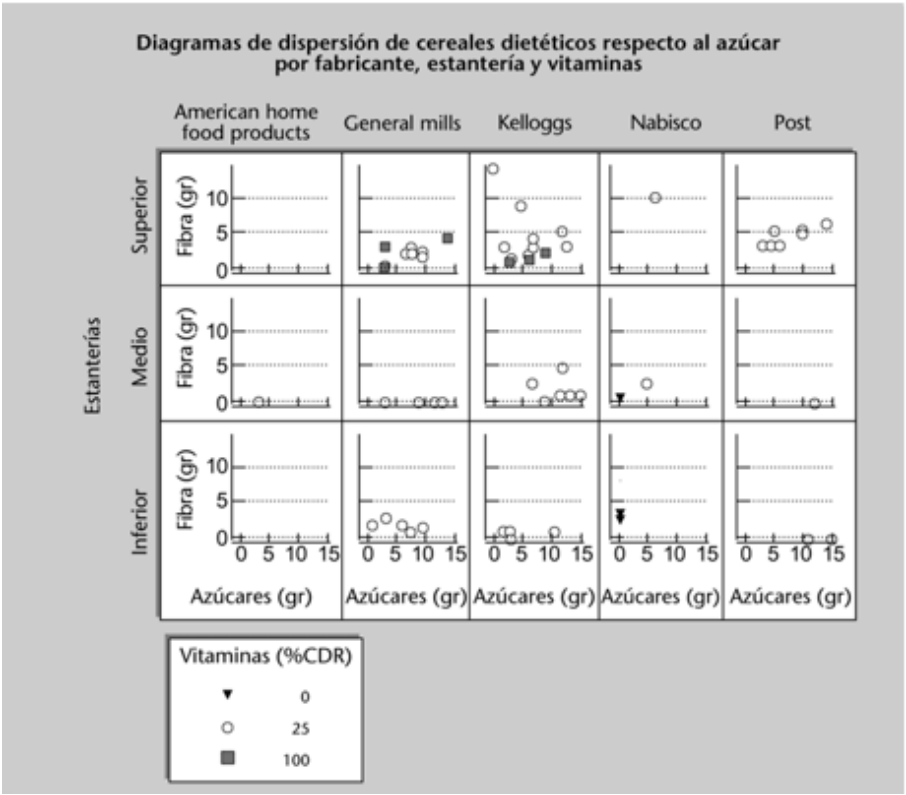
Es decir, si trabajamos con observaciones que tienen veinte atributos, conceptualmente estamos trabajando en un espacio de veinte dimensiones que

no podemos visualizar de ninguna manera. Ahora bien, proyectando parte de estos atributos en representaciones tridimensionales, o mejor bidimensionales, podemos extraer algún tipo de intuición que después podemos confirmar o refutar con otro tipo de herramientas –procedentes de la estadística y del aprendizaje automático–, y empezar un auténtico proceso de minería de datos.

El tratamiento de este problema admite varias formas. En general, se hacen combinaciones de histogramas o gráficos de dispersión para varias variables y distintas fuentes. Estas técnicas son bastante comunes y útiles para tratar de comparar rendimientos de centros dispersos geográficamente.

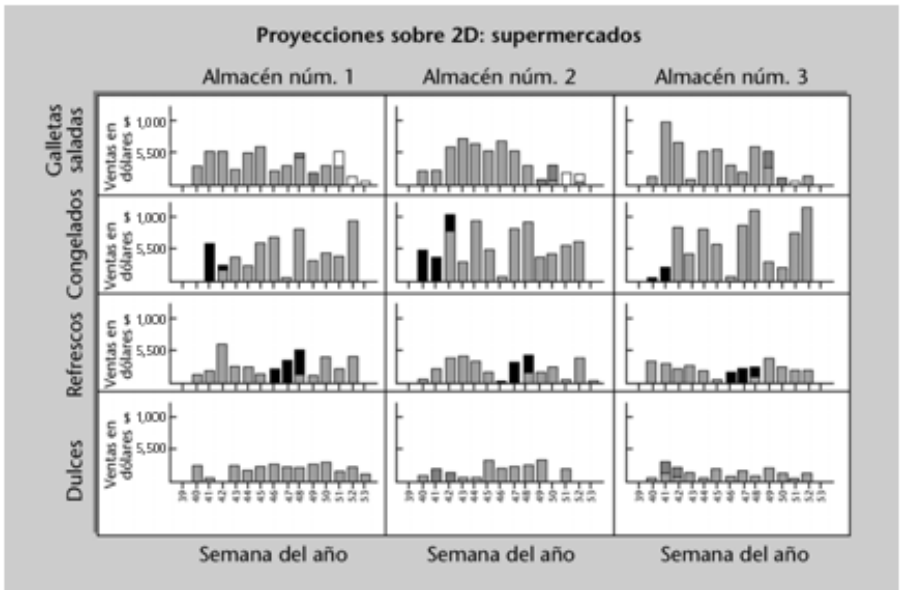
Veamos un ejemplo de técnicas de proyección sobre 2D: azúcar y disposición en las estanterías. Aquí tenemos un diagrama de dispersión que relaciona el contenido de fibra y azúcar para varias marcas de cereales en relación con la posición que ocupan dentro de las estanterías de un supermercado.

Figura 20. Diagramas de dispersión de cereales dietéticos respecto al azúcar por fabricante, estantería y vitaminas



El campo de herramientas de visualización de datos tiene una gran actividad y existen muchas herramientas para explorar datos. Aquí tenemos otro ejemplo de proyección sobre 2D:

Figura 21. Proyecciones sobre 2D: supermercados



En este ejemplo contamos con una tabla que presenta histogramas que relacionan la semana del año en la que se han recogido los datos con el nivel de ventas alcanzado por varios productos (caramelos, bebidas no alcohólicas, alimentos congelados y galletas saladas) para tres supermercados diferentes.

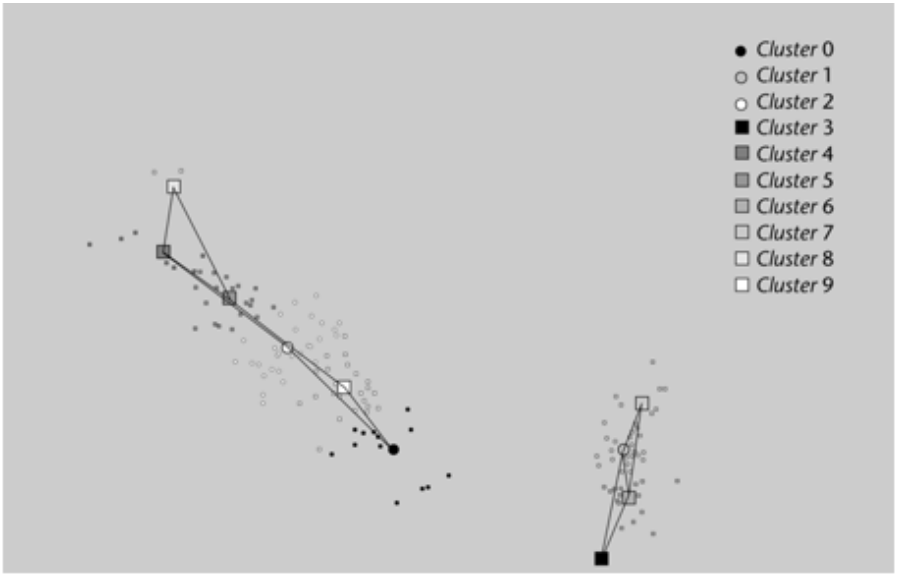
Otras herramientas permiten combinar varias formas de visualización en una sola. En la figura siguiente podemos ver cómo se combina la estructura de un árbol de decisión en tres dimensiones con los histogramas que reflejan la distribución de los valores de la partición que se induce a escala de nodo:

Figura 22. Visualización en 3D de histograma en la participación correspondiente a un nodo

Para el caso de agregación de datos, en inglés *clusters*, en el que se intenta encontrar grupos de datos parecidos, la representación de nubes de puntos sobre espacios bidimensionales permite estudiar cada grupo de objetos según las características elegidas.

Veamos un ejemplo de utilidad de proyecciones 2D en casos de agregación: en un caso de agregación interesa encontrar qué grupos de objetos están próximos entre sí. Por tanto, una ayuda importante para dicha tarea consiste en representar gráficamente el camino que conecta los grupos de objetos (*clusters*) más próximos o fuertemente relacionados:

Figura 23

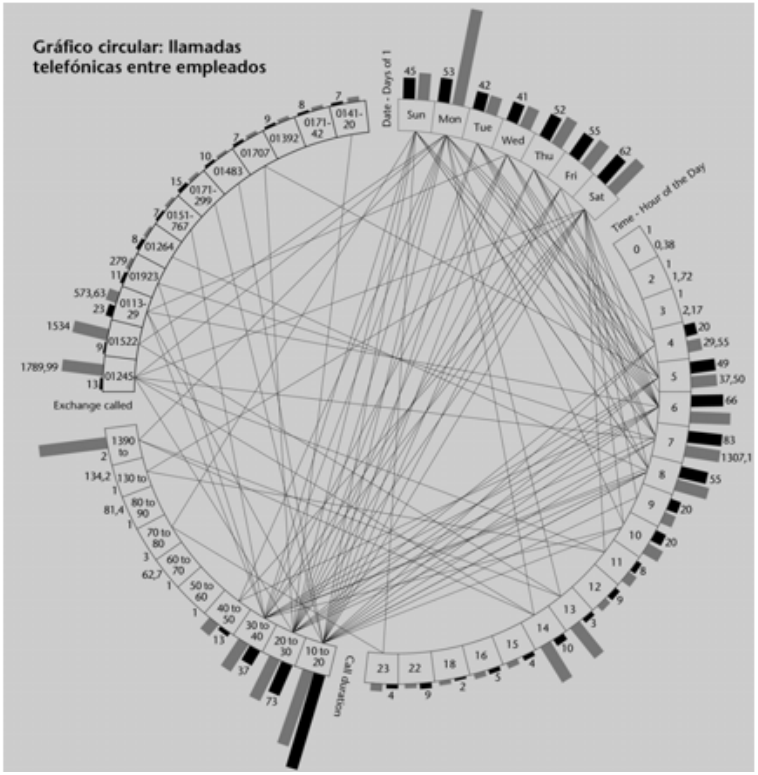


En tal caso, el gráfico indica que hay diez *clusters* y asigna un color diferente a los elementos de cada *cluster*. Si leemos los colores, podremos ver que hay fuertes relaciones (líneas continuas) entre los *clusters* 1, 3, 6 y 7, y entre los grupos 0, 2, 4, 5, 8 y 9.

El área de las técnicas de visualización es muy activa y cada vez ofrece más variantes de presentación, algunas bastante curiosas, como por ejemplo los gráficos circulares, en los que cada nodo representa algún tipo de variables.

Veamos finalmente un ejemplo de utilidad de gráfico circular. El que vemos aquí relaciona llamadas entre varios números de teléfono de los empleados de una empresa:

Figura 24. Gráfico circular: llamadas telefónicas entre empleados



A la vista del gráfico, ¿adiviná is quién trabaja con quién? Este tipo de visualizaciones permite extraer un conocimiento «rápido» del conjunto de datos que puede ser explotado después.

La potencia de las distintas herramientas de visualización se pone de manifiesto por su correcta conexión con el resto de las utilidades para selección y preparación de datos, como también con los métodos de minería de datos posteriores a la primera fase de intuición que aportan las claves visuales. Es importante tener en cuenta que estas visualizaciones mostradas como ejemplos no son óptimas desde el punto de vista de la visualización de datos, son simplemente una muestra de la capacidad exploratoria que tienen como herramientas de análisis visual.

3.2. *Data warehouse*

El concepto original de *data warehouse* fue presentado por William Inmon y comercializado por IBM con el término *information warehousing*, estableciendo la analogía entre los almacenes físicos de las empresas donde podían localizarse de forma flexible los materiales según la necesidad y el equivalente en cuanto a los datos de interés de sus distintas áreas de la empresa.

La intención de la propuesta de *data warehouse* es suministrar una infraestructura para tomar decisiones con cuatro objetivos fundamentales:

1) Regular el acceso a los sistemas de información y almacenamiento de datos según los diferentes tipos de usuarios y grupos de trabajo de manera más flexible y dinámica que las bases de datos tradicionales.

2) Facilitar la representación de datos y la reconfiguración de esa representación según las necesidades de toma de decisiones de la empresa, que cambian a medida que cambia el entorno competitivo.

3) Construir un modelo de datos corporativo que permita un mejor mantenimiento y evolución que los modelos actuales.

4) Mantener la independencia entre los procedimientos dirigidos a los usuarios finales y los de administración de datos, separando un tipo de procedimientos del otro.

El *data warehouse* puede considerarse como una manera de agrupar datos procedentes del sistema de transacciones de la empresa con los datos que son necesarios para el trabajo diario de grupos situados en jerarquías intermedias y los decisores de alto nivel. Ni que decir tiene que cada tipo de entorno posee requisitos diferentes y maneras de ver los datos también distintas y cambiantes. El *data warehousing* agrupa varios tipos de herramientas y tecnologías.

Algunas de las tecnologías que, sin ser nuevas, son utilizadas o tienen relevancia con el *data warehousing* son las siguientes:

1) Sistemas de gestión de bases datos que soporten proceso paralelo.

2) Herramientas de conversión automática de datos.

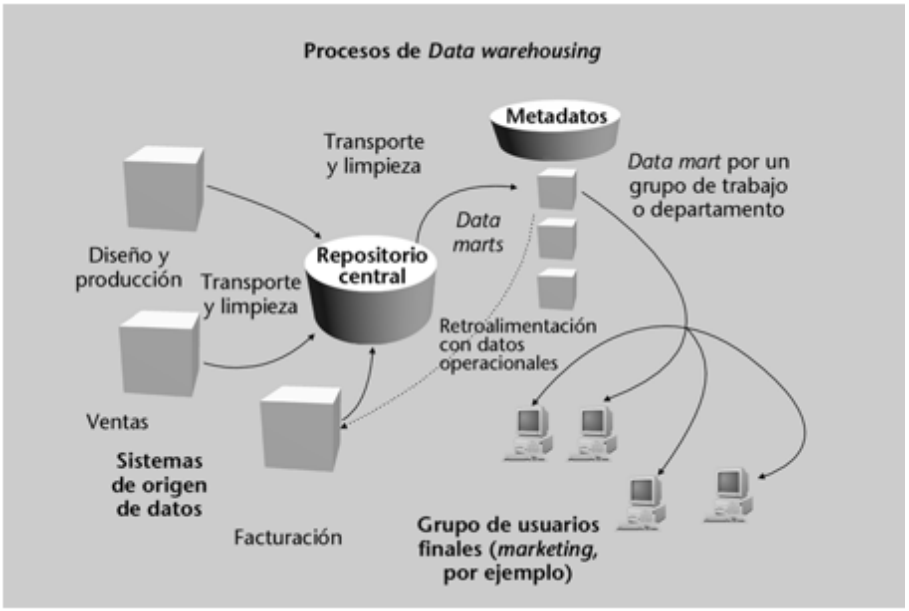
3) Tecnologías cliente/servidor para acceder a datos distribuidos en plataformas diferentes.

4) Integración de herramientas de análisis y relación con sistemas de toma de decisiones, sistemas de toma de decisiones en grupo y sistemas de información para ejecutivos.

El aspecto más crítico del *data warehousing* es probablemente el modelado de los distintos usos y perspectivas que se tiene de los datos, así como la integración transparente de los distintos productos de software y su actualización y mejora continuas y lo más automáticas posible.

Los procesos de *data warehousing* hacen un uso extensivo de grandes volúmenes de datos, en particular datos históricos (cinco a diez años) que se manipulan a varios niveles para analizar datos, sintetizar otros nuevos o ponerlos en relación con los factores críticos de éxito de la empresa.

Figura 25. Procesos de *data warehousing*



Un sistema de *data warehousing* contiene, normalmente, los componentes siguientes:

1) Sistemas de origen de datos. Sistemas que recogen los datos en el nivel más bajo (en el sentido de que recogen los datos con un mínimo de abstracción). Por ejemplo, los que recogen los datos procedentes de puntos de venta. La tarea del sistema de *data warehousing* consiste en poder integrar las informaciones procedentes de distintas fuentes de manera coherente y aportar una descripción un poco más elevada.

2) Transporte y limpieza de datos. Sistemas de software que se encargan de «limpiar» los datos, en el sentido que ya hemos explicado, y llevarlos a otros enclaves donde se guardarán en la forma o formas adecuadas. Tradicionalmente, este tipo de procedimiento era tarea de programación y resultaba difícilmente ampliable. Los productos de transporte y limpieza disponibles hoy día adoptan una óptica más de especificación, en la que se indica de dónde proceden los datos y qué les tiene que suceder sin llegar a procesarlos. Por norma general, esta parte implica una descripción de los datos en otro lenguaje de descripción de datos, y crea lo que se conoce como *metadatos*.

3) Repositorio central. Lugar principal donde se guardan los datos del almacén. Consta de los elementos siguientes:

a) Hardware ampliable. La ampliabilidad del hardware radica en el hecho de que permite aumentar sin demasiada perturbación tanto la rapidez de cálculo (computación paralela) como el volumen de datos (en torno a los terabytes).

b) Sistema de bases de datos relacional. Las bases de datos relacionales del repositorio central están especialmente pensadas para mejorar la construc-

ción dinámica de índices, las operaciones de copia y mantenimiento, y el procesamiento de consultas variadas y no estáticas en el tiempo.

c) Modelo lógico de datos. Finalmente, el modelo lógico de datos tiene como objeto la intercambiabilidad de datos entre los distintos componentes de la empresa y la mantenibilidad del repositorio.

4) Metadatos. Como hemos dicho, son «datos sobre los datos» que introducen un grado de abstracción más elevado con respecto a los componentes básicos, que son las tablas y las relaciones. Hay una gran variedad de componentes de los metadatos, que, además de facilitar la comprensión y la administración de los datos a los administradores del *data warehouse*, intentan mejorar la comprensión y el acceso por parte de los usuarios finales.

5) *Data marts*. Se trata de «personalizar» la visión, los componentes y los contenidos del *data warehouse* según las necesidades de los distintos grupos de trabajo. Los datos de una vista combinan los de varias tablas relacionales, probablemente distribuidas.

6) Herramientas de realimentación operativa. Recogen los datos procedentes de los sistemas de toma de decisiones integrándolos en el repositorio central. Esta es una desviación notable respecto del uso tradicional de las herramientas de toma de decisiones operacional. Por ejemplo, integran criterios para hacer pedidos a proveedores en relación con niveles de *stock* y grado de cumplimiento del proveedor en cuestión, o integran ayudas para trabajar con clientes. Este aspecto del *data warehouse* permite, por ejemplo, ofrecer sugerencias a un cliente después de que haya contestado a una serie de preguntas directamente al personal de atención al cliente. Es uno de los aspectos en los que las herramientas de minería de datos ofrecen más resultados.

La relación entre *data warehousing* y minería de datos es considerada por algunos como inclusiva, en el sentido de que las herramientas de minería de datos forman parten del entorno de *data warehousing*.

En el concepto *data warehousing* pueden agruparse plataformas de software, herramientas de extracción y conversión de datos, bases de datos preparadas para consultas complejas y dinámicas, herramientas de análisis de datos y herramientas de gestión de bases de datos.

Los *data warehouse* constan de herramientas para la mejora de la comprensión y el acceso por parte del usuario como anotaciones en el modelo lógico, mapeo del modelo lógico en los sistemas fuente de datos, vistas y fórmulas más comunes para acceder a los datos e información de seguridad y acceso.

Lectura recomendada

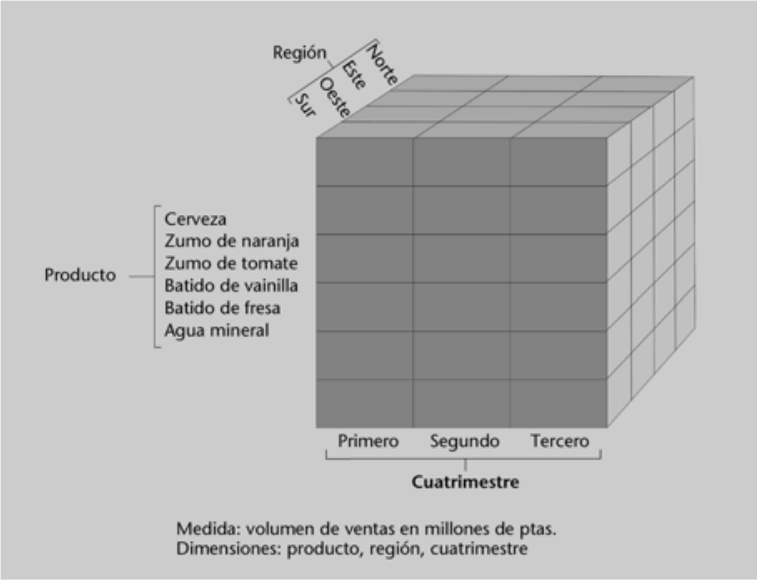
Encontraréis información acerca del concepto original de *data warehouse* en la obra siguiente:
W. Inmon (1996). *Building the Data Warehouse* (2.ª ed.). Nueva York: John Wiley & Sons.
Y un trabajo más reciente del mismo autor sería el siguiente:
W. Inmon; D. Strauss; G. Neushloss (2010). *DW 2.0: The architecture for the next generation of data warehousing*. Elsevier.

3.3. Métodos OLAP

OLAP es la sigla de la expresión inglesa *on-line analytical processing*. Los métodos OLAP aparecieron para analizar los datos de ventas y marketing, pero también para procesar datos administrativos y consolidar datos procedentes de varias fuentes de cara a efectuar un análisis de rentabilidad, mantenimiento de calidad y otros tipos de aplicaciones que se caracterizan por que redefinen de manera continua y flexible el tipo de información que hay que extraer, analizar y sintetizar (en comparación con las bases de datos tradicionales, dirigidas a responder consultas bastante prefijadas y rutinarias).

Los sistemas OLAP se alimentan de los datos generados por los sistemas transaccionales (facturación, ventas, producción, etc.). Herramientas típicas de OLAP son las que permiten realizar un análisis multidimensional de los datos en contra de las típicas facilidades de creación de resúmenes e informes propios de los sistemas de bases de datos tradicionales.

Figura 26



La unidad de datos de OLAP es el «cubo», una representación de los datos que permite «cortarlos» a rebanadas y verlos desde las perspectivas de muchos grupos diferentes de usuarios. La característica principal de los cubos es que optimizan las consultas. Normalmente, se guardan en forma de tabla relacional especial que facilita ciertos tipos de consultas. Por ejemplo, hay columnas de las tablas que se denominan *columnas de dimensión*, y que facilitan y prevén datos para resúmenes e informes. Las columnas llamadas *columnas agregadas* permiten precalcular cantidades como recuentos, sumas y medias.

Construir un cubo requiere un análisis detallado de las necesidades de datos del grupo de usuarios al que va dirigido, y puede requerir, asimismo, bastante tiempo, tanto de diseño como de instalación por primera vez. Compensa por el hecho de que facilita extraordinariamente las tareas de análisis de datos de

los distintos grupos de usuarios y, una vez establecido, resulta más sencillo de modificar que las tablas relacionales tradicionales.

3.4. Sistemas OLTP

OLTP es la sigla de la expresión inglesa *on-line transactional processing*. Los sistemas de procesamiento de transacciones en línea (OLTP) tienen como objetivo guardar la integridad de los datos necesarios para administrar una organización de manera eficiente.

Así pues, los sistemas OLTP buscan mantener modelos de datos que correspondan a la visión que cada empleado (o tipo de empleado) tiene de la organización. En lugar de ver la organización como una estructura de datos organizada de tablas y relaciones, las herramientas de OLTP la presentan en forma de jerarquías y dimensiones, de manera que podemos observar los mismos datos desde perspectivas diferentes.

Los sistemas tradicionales son dinámicos, en la medida en que siempre están siendo actualizados con nuevos datos. Para analizarlos es necesario hacer una «fotografía» de su estado en un momento dado y aplicar las herramientas de análisis correspondientes. Llevar a cabo este tipo de trabajo solo con las operaciones de consulta propias de las bases de datos tradicionales no es fácil, y puede inducir a una degradación del rendimiento general del sistema. Asimismo, el sistema de bases de datos puede no estar preparado para guardar el resultado de estos análisis.

En cambio, los sistemas de OLTP (como los de OLAP, en cierto modo) permiten descargar el sistema central (ocupado, quizá, en procesos transaccionales) y efectuar este tipo de operación al mismo tiempo que permite guardar sus resultados. Las herramientas de minería de datos pueden dar algún servicio a este tipo de análisis.

3.5. Estadística

La tarea consistente en analizar grandes volúmenes de datos ha sido y sigue siendo el reinado de la estadística, en concreto, el análisis de datos. El enfoque tradicional de la estadística se dirige a la recopilación de datos adecuada para la interpretación, en particular a la inferencia de características de una población a partir de las muestras recogidas.

La idea de minería de datos ha situado las técnicas estadísticas clásicas ante una gran oportunidad práctica y también ante la necesidad de crear herramientas que, aun manteniendo la sólida fundamentación teórica aportada por esta disciplina, den respuestas fácilmente comprensibles a usuarios no

siempre bien preparados estadísticamente dentro de los límites de tiempo impuesto por la velocidad que requieren los nuevos entornos de trabajo.

3.6. Aprendizaje automático

El aprendizaje automático es aquella parte de la inteligencia artificial que estudia cómo los sistemas inteligentes son capaces de desarrollar conocimientos y habilidades nuevas a partir de su experiencia.

En concreto, los métodos de aprendizaje inductivo (Michell, 1985) buscan la extracción de conceptos, pautas de conducta y planes nuevos; en general, conocimientos nuevos a partir de la observación de los datos del entorno o del propio comportamiento del sistema inteligente. La plétora de métodos aportados desde este campo y su insistencia en primar la expresión simbólica y no numérica del conocimiento también han convertido sus métodos en relevantes para la tarea de minería de datos.

Resumen

Podemos decir, a modo de resumen, que la minería de datos integra resultados de disciplinas como son las bases de datos (con su extensión a *data warehousing*, OLAP y OLTP), la estadística, el aprendizaje automático y la visualización. Hemos de señalar que la propia propuesta de minería de datos ha generado una gran actividad en todos estos campos, que se han visto obligados a modificar algunos de los supuestos para llegar a proporcionar la calidad de resultado exigida por los nuevos objetivos.

En minería de datos se busca la obtención de conocimiento nuevo, válido y útil para los objetivos que se plantee quien emprenda dicho proceso. El resultado de un proceso de minería de datos es un modelo que tiene que ser lo más comprensible posible. Es importante que se pueda interactuar con este proceso y aprovechar el conocimiento *a priori* del que se disponga.

Los procesos de minería de datos se basan en resultados procedentes de la investigación y el desarrollo en bases de datos, estadística, aprendizaje automático y visualización. Debemos entender la minería de datos como un proceso continuo que integra los aspectos siguientes:

1) Definición del objetivo del proyecto de minería de datos, precisando la tarea principal que hay que realizar y eligiendo el método más adecuado según las circunstancias.

2) Selección de los datos relevantes.

3) Preparación de los datos de cara a asegurar que sean válidos y se encuentren en condiciones de ser utilizados por el método seleccionado.

4) Minería de datos propiamente dicha, es decir, aplicación sobre los datos ya preparados del método elegido y construcción del modelo correspondiente.

5) Interpretación del modelo obtenido, que puede provocar la revisión de algunas de las fases anteriores.

6) Integración en el sistema de tratamiento de información, que comprende la observación del rendimiento y, en caso de cambio del entorno o «envejecimiento» del modelo, inicio de un proceso de minería de datos nuevo.