Rodolfo Fernández González

Doctor en Filosofía, fue profesor titular de Lógica e Inteligencia Artificial de la Facultad de Psicología de la Universidad Complutense de Madrid.

Se formó en Informática en HP y UNISYS, colaborando como programador y analista en Investigación y Mercado, desde 1972 a 1975, y como responsable de Enseñanza Asistida por Ordenador y de diversos proyectos en el INCIE (Instituto Nacional de Ciencias de la Educación) desde esa fecha hasta 1980.

Colabora en Indra como Gestor de Proyectos desde 1988 hasta su muerte en 1999.

Técnica de Inteligencia Artificial en Minería de Datos

Rodolfo Fernández González
Ingeniería del Conocimiento

Resumen

Se describen las distintas técnicas que dentro del ámbito de la Inteligencia Artificial se han usado con éxito, junto a otras de tipo estadístico, en lo que se ha dado en llamar «minería de datos». Para cada una, se referencian un conjunto de productos comerciales que las soportan.

El concepto de **Data Warehouse** implica, obviamente, no sólo la reorganización, y puesta a disposición de los usuarios de los datos disponibles en la compañía, sino, sobre todo, su explotación para la toma de decisiones empresariales. Este proceso puede requerir la solución de problemas que se clasifican en los siguientes tipos:

- **Clasificación de casos**. Es la aplicación más extendida. Puede tratarse de una segmentación de clientes anteriores, con vistas a posteriores acciones de marketing sobre ellos, o de un estudio de su comportamiento con vistas a poder predecir en el futuro cuál será el comportamiento de nuevos clientes.
- **Derivación de dependencias:** Se trata de construir explicaciones causales a partir de un conjunto de datos (por ejemplo, siniestros de automóviles), con el fin de utilizar dichas explicaciones para la predicción (por ejemplo, del nivel de riesgo de nuevos solicitantes de seguros)
- **Detección de desviaciones:** Se trata de identificar casos anómalos a partir de la experiencia anterior. Por ejemplo, revisión de facturas de usuarios en *utilities.*
- En todos los casos, al exponer al sistema a la información disponible en la compañía, lo que se espera es que *«aprenda»* de esos datos acumulados, y que utilice el conocimiento aprendido en la resolución de nuevos problemas. Como veremos a continuación, sólo algunas de las tecnologías disponibles -induccion de reglas, generación de casos- hacen explícito -y utilizable con otros propósitos- el conocimiento aprendido.
- Cualquiera que sea la solución de base adoptada para el *Data Warehousing* (SGBD relacional, CMS[1], etc.), el objetivo final consiste en alimentar el EIS con la información relevante para cada perfil de usuario, orientada a sus objetivos y a su modo de trabajo.

[1] Productos relacionales híbridos, orientados a objetos, que pueden manejar datos tanto estructurados como no estructurados.

Para la obtención de dicha información, se requiere la ejecución de tareas en tres niveles distintos:

1. Construcción del mapa de metadatos, que supone la cualificación de las diversas fuentes de datos existentes respecto a los objetivos de cada usuario.
2. Extracción de la información relevante de las fuentes de datos seleccionadas
3. Presentación de dicha información al usuario de acuerdo con las especificaciones recogidas respecto a su modo de trabajo.

En lo que sigue vamos a referirnos a algunas técnicas especiales aplicables sobre todo, en el segundo nivel (extracción). Estas técnicas se han desarrollado dentro del ámbito de la Inteligencia Artificial, y actualmente están siendo utilizadas de forma conjunta con técnicas bien conocidas de análisis estadístico *(clustering,* análisis dimensional, etc.), bajo el nombre de *data mining* o *knowledge discovery.* Las principales técnicas aplicables son:

1. **Redes neuronales,** especialmente mapas de Kohonen y redes de *backpropagation* o perceptrones multicapa.
2. Generadores de reglas por **inducción,** explotables mediante motores inferenciales.
3. Generadores de **bases de casos** a las que se accede mediante técnicas CBR.
4. Sistemas clásicos de heurísticos.

Las estrategias de aplicación de estas técnicas adoptan actualmente varias formas distintas:

1. Algunas de ellas se encuentran disponibles actualmente en algunas herramientas *standalone* de minería de datos, como IDIS (Information Discovery System) de IntelligenceWare y Database Mining Workstation, de HNC Software, que facilitan la «excavación» de vastas cantidades de información.
2. En otros casos, se trata de desarrollos que integran como subsistemas -por ejemplo, vía APIs- herramientas de desarrollo específicas (como Rules de ILOG o CasePoint Search Engine de Inference). Neural Works, de NeuralWare, permite convertir una red entrenada en una función C. Si se desea reentrenar a la red, se requiere el uso adicional del Designer Pack.
3. En tercer lugar, puede optarse por implementar directamente los algoritmos correspondientes en desarrollos convencionales que se alimentan de los datos proporcionados por el entorno de Data Warehouse.
4. Recientemente, se está prestando gran atención, especialmente en el terreno de la minería de datos, al nuevo paradigma de los agentes inteligentes. Un agente inteligente no es, en realidad, una nueva tecnología, sino una metáfora para describir un tipo específico de funcionalidad de sistema que exhibe las siguientes características[2] A:

- Autonomía: puede trabajar en ausencia del usuario y puede adaptar sus acciones a condiciones cambiantes del entorno.
- Personalización: Mantiene y explota información acerca de ciertos intereses o hábitos de un usuario o grupo de usuarios.
- Multiaplicación/Multiplataforma. .Capacidad de aprendizaje: Cambia su conducta en el tiempo depen- diendo de las acciones o del *feedback* del usuario.
- Capacidad de cooperación con otros agentes.

En las secciones siguientes se ofrece una breve revisión de las herramientas de minería de datos y descubrimiento de conocimiento actualmente disponibles3, que incluyen una o más de estas técnicas.

[2] Ver What's in a Name: Intelligent Agents, Gartner Group, ATG, Research Note, May 23, 1994; Intelligent Agents: A Day in the Life of a Software Agent, Gartner Group, ATA, Strategic Analysis Report, September 28, 1995.

Redes neuronales para la minería de datos

Las redes neuronales[3] se han revelado como un útil instrumento para obtener información a partir de grandes masas de datos. En la minería de datos se han utilizado, básicamente, dos tipos distintos de redes neuronales: redes de Kohonen y perceptrones multicapa o redes de *backpropagation.*

Las redes de Kohonen o mapas autoorganizativos son redes neuronales que forman un mapa bidimensional de rasgos a partir de los datos de entrada de forma que cada caso queda agrupado en clases o *clusters* de máxima similaridad. El rasgo diferencial más importante de este tipo de red es que aprende sin supervisión, es decir, es capaz de descubrir clases de casos. Normalmente, se combina con otras redes de aprendizaje supervisado, como los perceptrones multicapa o redes de *backpropagatíon.* Estos sistemas de aprendizaje supervisado permiten identificar clases no linealmente separables reajustando -a partir del error de salida- los pesos de las conexiones del nivel anterior de la red, lo que se repite hasta llegar a la capa de entrada. Puesto que esta técnica ya ha sido objeto, con otros propósitos, de la atención del **Bit[4],** no nos extenderemos más sobre ellas.

Entre las herramientas de redes neuronales actualmente disponibles se encuentran NeuralWorks, de NeuralWare; Darwin, de Thinking Machine; Recon Data Mining System, de Lockeed Martin Product and Services; Neural Connection, de SPSS Inc; MATLAB Neural Network Toolbox, de The Math Works; AIM y AIM Statnet, de Abtech Corp.

Clasificación por inducción

Una importante familia de técnicas es la que se agrupa en torno a los llamados «métodos inductivos». Entre estos podemos señalar:

- Arboles, grafos y tablas de decisión.
- Medidas de proximidad *(nearest-neighbors,* CBR).
- Redes bayesianas, *belivef networks.*

Puesto que ya se ha recogido anteriormente en Bit[5] la técnica de razonamiento basado en casos, que utiliza técnicas de medida de proximidad para recuperar información almacenada bajo la forma de casos, vamos a examinar brevemente en lo que sigue lo más fundamental de los mecanismos de inducción.

Los algoritmos de clasificación por inducción nos permiten obtener resultados de un proceso de aprendizaje supervisado. El sistema de inducción se alimenta con una colección de casos o ejemplos *(trainig set)* extraídos del almacén de datos. Cada caso se describe mediante:

- Un conjunto de atributos
- Su pertenencia o no a una clase determinada. Al tratarse de aprendizaje supervisado, se identifica la clase a la que pertenece el caso. Por ejemplo, si se trata de caso de fraude, de una factura incorrecta, de un cliente moroso, etc.

Obviamente, no todos los atributos suelen ser relevantes para la clasificación y en la elección de los atributos relevantes intervienen los expertos de la compañía. Esa simplificación no reduce, sin embargo, el número de casos a considerar, que, en las situaciones en las que el *data mining* es aconsejable, produciría una explosión combinatoria intratable con otros métodos. Si tenemos, por ejemplo, 12 atributos por caso, y 5 valores por atributo (con una frecuencia parecida) tendríamos $5^{12} = 2.44 \times 10^8$ combinaciones. Si en este caso tuviéramos un millón de casos observados, sólo habríamos examinado el 0.4% del total. De esta diversidad prácticamente inmanejable es de la que se hacen cargo estas técnicas inductivas.

---

[3] Una fuente de información en Internet periódicamente actualizada es http://info.gte.com/-kdd/ La última actualización para esta nota es de Septiembre de 1996. Sólo se recogen aquí las herramientas disponibles comercialmente. Existen, además, multitud de librerías de dominio público, y herramientas de investigación.

[4] «Redes Neuronales», **BIT** nº 1, 1994

[5] "Centros de Soporte a usuarios *y/o* clientes (Help-desk), **BIT** nº 9, Octubre 1995

Existen muchos algoritmos de inducción. Los más utilizados son el ID3 y el C4, de Quinlan (1979, 1983, 1986,1987,...), Y los AQ, AQ11, AQ15,...de Michalski (1969, 1983, 1986, ...). Cada uno de los algoritmos forma el núcleo de sistemas clasificatorios distintos, en los que se pueden utilizar árboles de identificación o de decisión, con rangos de valores discretos o continuos, y que pueden operar en modo *batch* o incremental.
El problema fundamental con el que se enfrenta cualquier sistema de clasificación, y también los sistemas de inducción, es el de la eliminación del ruido en el *training set* (ver fig.1). Para ello, se construyen árboles de **identificación** que son árboles de decisión en los que:

- Cada nodo no terminal está etiquetado con un atributo.
- Cada rama que sale de un nodo está etiquetada con un valor de ese atributo.
- Cada nodo terminal está etiquetado con un conjunto de casos, cada uno de los cuales satisface todos los valores de atributos que etiquetan el camino desde ese nodo al nodo inicial.

La aplicación de un atributo como criterio de selección clasifica los casos en distintos conjuntos (tantos como valores discretos del atributo). Se trata de construir el árbol de identificación más simple que sea consistente con el *training set.* Para ello hay que ordenar los atributos relevantes, desde la raíz a los nodos terminales, de mayor a menor fuerza clasificatoria. La «fuerza clasificatoria» de un atributo es su capacidad para generar particiones del *training set* que se ajusten en un grado dado a las distintas clases posibles, introduciendo de esta forma un orden en dicho conjunto. Sabemos por la teoría matemática de la información que ese orden (y el desorden o «ruido») de un conjunto de datos es medible. Lo que hacemos es medir la fuerza clasificatoria de un atributo mediante su capacidad para reducir la incertidumbre o «entropía». Construiremos, por tanto, el árbol de identificación siguiendo las siguientes fases:

1. Cálculo de la entropía que puede reducir cada atributo.
2. Ordenación de los atributos de mayor a menor capacidad de reducción de entropía.
3. Construcción del árbol de identificación siguiendo la lista ordenada de atributos.

Como es bien sabido, la medida de la entropía *E* viene dada por la fórmula:

$$E = \sum_{I=1}^{c} -P(c) \cdot \log_2 P(c)$$

donde, en este caso, *c* es el número de clases, y

$$P(c) = \frac{n_{rc}}{n_r}$$

esto es, la probabilidad de que un caso determinado pertenezca a la clase c, siendo n el número de casos en la rama *r* de la clase c, y n_r el número de casos en una rama *r.* Evidentemente, si un conjunto de casos contiene miembros de dos clases, y el número de elementos de cada clase es el mismo, el valor dado por la fórmula es igual a 1, el máximo posible.

Caso	$Atributo_1$	...	$Atributo_1$	Clase
$Caso_1$	$Valor_{11}$	...	$Valor_{1n}$	$Clase_1$
...	...	...	...	...
$Caso_m$	$Valor_{m1}$	...	$Valor_{mn}$	$Clase_k$

Fig. 1. Formato del *Training Seto*

Si sólo hubiera elementos de una clase, el valor de la entropía sería 0, el mínimo posible. A medida que nos movemos entre una distribución perfectamente equilibrada y una homogeneidad perfecta, la entropía va variando entre 0 y 1.
Una vez calculada la entropía de cada atributo, se establece una lista de los mismos ordenada de mayor a menor entropía. A continuación se puede proceder a la construcción del árbol, en las siguientes fases, hasta que cada nodo terminal contenga un subconjunto del *training set* lo más homogéneo posible:

1. Se selecciona un nodo terminal con un conjunto no homogéneo.
2. Se sustituye el nodo por una prueba sobre el primer atributo de la lista no utilizado todavía en la rama. Esta prueba dividirá el conjunto en subconjuntos mínimamente no homogéneos.
3. Sobre cada uno de los conjuntos resultantes se reitera el procedimiento, hasta agotar los atributos disponibles.

Entre las herramientas disponibles que utilizan directamente la estrategia de árboles de decisión se encuentran AC2, de Isoft; INO, de NASA COSMIC; Knowledge-SEEKER, de ANGOSS Software y SPSS CHAIO, de SPSS INC.
Los resultados de la inducción pueden tratarse ulteriormente de varias formas. Las dos más habituales son la obtención de casos destinados a integrarse en una base de casos que posteriormente se explotará con técnicas de CBR, o la obtención de reglas. Cada una de las reglas de clasificación que se obtienen a la salida tienen la siguiente forma:

- Antecedente: Conjunto de pares <atributo, valor>
- Consecuente: Identificación de una clase.

Para transformar el árbol de identificación en un conjunto de reglas:

- Se recorre cada rama de la red desde la raíz hasta el nodo terminal.
- El antecedente de la regla es la conjunción de los pares <atributo, valor> recogidos en cada nodo.
- El consecuente de la regla es el nodo terminal.

Normalmente, es necesario simplificar las reglas así obtenidas, lo cual se lleva a cabo eliminando antecedentes innecesarios, eliminando reglas innecesarias y reduciendo todas las reglas con el mismo consecuente a una sola regla, que puede ser una regla por defecto.
El formato de regla sólo resulta útil si al conjunto de reglas obtenido se le añaden reglas heurísticas obtenidas de los expertos. En este caso, las reglas obtenidas pueden aplicarse ulteriormente para clasificar nuevos casos utilizando un motor inferencial clásico. Si no hay heurísticos, resulta más aconsejable utilizar el conocimiento conocido bajo la forma de base de casos. En la fig. 2 puede verse un esquema de las etapas implicadas en la generación tanto de sistemas CBR como de sistemas de reglas, partiendo de un proceso de inducción.
Entre las herramientas actualmente disponibles que permiten la obtención de reglas a partir de árboles de decisión/identificación, además de las ya mencionadas Darwin, de Thinking Machine, y Recon Data Mining System, de Lockeed Martin Product and Services, se encuentran Datalogic, de Reduct Systems, e IDIS, de Information Discovery. La explotación ulterior de dichas reglas -con heurísticos añadidos- puede llevarse a cabo con herramientas como Rules, de ILaG, o Rete++ de Haley Enterprises (librerías C++ que

incorporan un motor inferencial y un completo sistema de gestión de agenda). Como bases de casos, pueden explotarse con las herramien- tas de Inference Corp. (CasePoint, CBR Express, Generator, Tester y Search Engine), distribuídas por Indra.

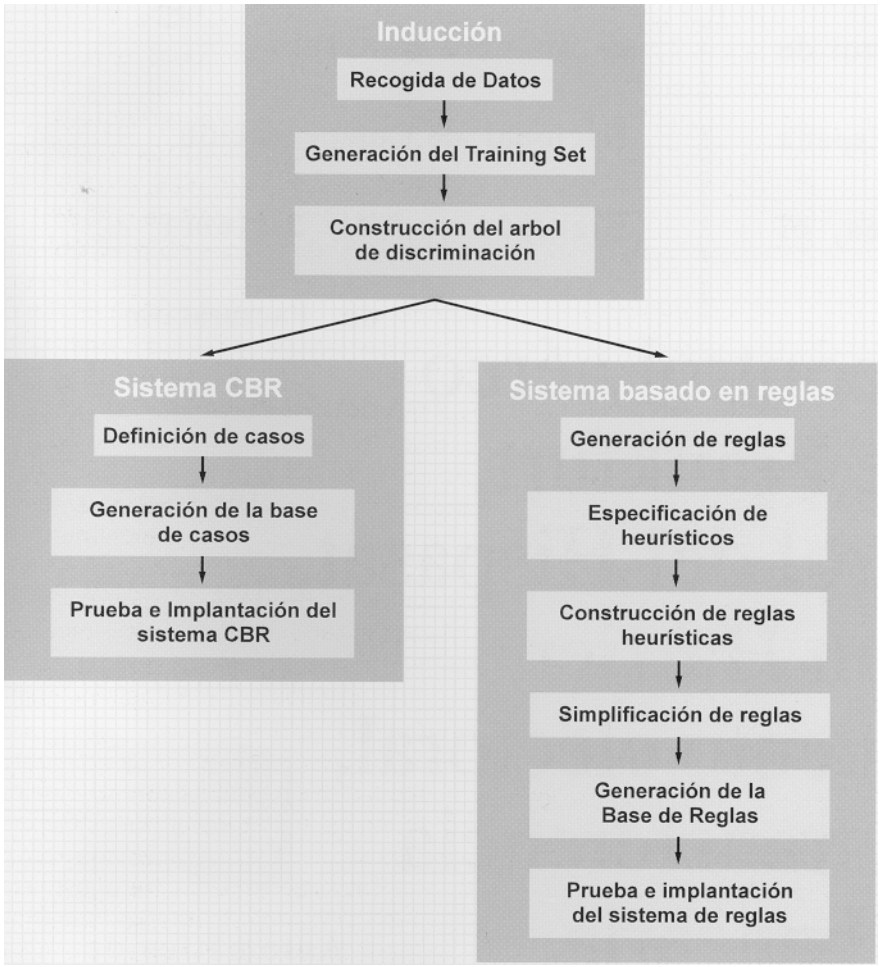

Fig. 2. Etapas del proceso de construcción de sistemas CBR y de reglas basados en inducción.