Capítulo 1

Introducción

Hasta hace no demasiado tiempo se utilizaba el término "procesamiento de datos" para describir la utilización de los ordenadores en distintos ámbitos. Hoy se utiliza otro término, IT [Information Technology], que se refiere a lo mismo pero implica un cambio de enfoque. Se hace énfasis no únicamente en el procesamiento de grandes cantidades de datos, sino en la extracción de información significativa de esos datos.

Los datos son información cruda, colecciones de hechos que deben ser procesados para que sean significativos. La información se obtiene asociando hechos (en un contexto determinado). El conocimiento utiliza la información obtenida en un contexto concreto y la asocia con más información obtenida en un contexto diferente. Finalmente, la sabiduría (término que nadie utiliza en IA) aparece cuando se obtienen principios generales de fragmentos de conocimiento.

Hasta ahora, la mayor parte del software ha sido desarrollada para procesar datos, información a lo sumo. En el futuro se trabajará con sistemas que procesen conocimiento. La clave reside en asociar elementos de información provenientes de distintas fuentes y sin conexión obvia de tal forma que la combinación nos proporcione beneficios. Este es uno de los desafíos más importantes de la actualidad: la construcción de sistemas que extraigan conocimiento de los datos de forma que sea práctico y beneficioso.

ROGER S. PRESSMAN
Ingeniería del Software: Un Enfoque Práctico

El aprendizaje en Inteligencia Artificial [53] [92] [94] [140] se entiende como un proceso por el cual un ordenador acrecienta su conocimiento y mejora su habilidad. En él se resaltan dos aspectos complementarios: el refinamiento de la habilidad y la adquisición de conocimiento. Tal como lo definió Simon, el aprendizaje denota cambios en el sistema que son adaptativos en el sentido de que permiten al sistema hacer la misma tarea a partir de la misma posición de un modo más efectivo.

Muchas de las técnicas de aprendizaje usadas en IA están basadas en el aprendizaje realizado por los seres vivos. Para ellos, la experiencia es muy importante, ya que les permite no volver a cometer los mismos errores una y otra vez. Además, la capacidad de adaptarse a nuevas situaciones y resolver nuevos problemas es una característica fundamental de los seres inteligentes. Por lo tanto, podemos aducir varias razones de peso para estudiar el aprendizaje: en primer lugar, como método de comprensión del proceso de aprendizaje (desde el punto de vista de la Psicología) y, en segundo término, aunque no por ello sea menos interesante, para conseguir programas que aprendan (desde una perspectiva más propia de la Inteligencia Artificial).

Una primera clasificación de las técnicas de aprendizaje existentes se puede realizar atendiendo a la filosofía seguida en el proceso de adquisición del conocimiento:

- En el *aprendizaje supervisado* (o aprendizaje a partir de ejemplos, con profesor), los ejemplos de entrada van acompañados de una clase o salida correcta. Esta familia de técnicas engloba al aprendizaje memorístico [*rote learning*], a los modelos de aprendizaje por ajuste de parámetros y a una amplia gama de métodos de construcción de distintos tipos de modelos de clasificación, desde árboles de decisión hasta listas de decisión.

- En el *aprendizaje no supervisado* (aprendizaje por observación, sin profesor) se construyen descripciones, hipótesis o teorías a partir de un conjunto de hechos u observaciones sin que exista una clasificación a priori de los ejemplos. Este tipo de aprendizaje es el que realizan los métodos de agrupamiento o *clustering*.

El aprendizaje con profesor o aprendizaje supervisado, también conocido como clasificación, es uno de los problemas más estudiados en Inteligencia Artificial. En particular, el objetivo de cualquier algoritmo de aprendizaje supervisado es construir un modelo de clasificación a partir de un conjunto de datos de entrada, denominado conjunto de entrenamiento, que contiene algunos ejemplos de cada una de las clases que pretendemos modelar. Los casos del conjunto de entrenamiento incluyen, además de la clase a la que corresponden, una serie de atributos o características que se utilizarán para construir un modelo abstracto de clasificación. El objetivo del aprendizaje supervisado es la obtención de una descripción precisa para cada clase utilizando los atributos incluidos en el conjunto de entrenamiento. El modelo que se obtiene durante el proceso de aprendizaje puede utilizarse para clasificar nuevos ejemplos (casos cuyas clases se desconozcan) o, simplemente, para comprender mejor los datos de los que disponemos. Formalmente, un modelo de clasificación se puede definir de la siguiente manera [39]:

> Si suponemos que todos los ejemplos que el modelo construido ha de reconocer son elementos potenciales de J clases distintas denotadas ω_j, llamaremos $\Omega = \{\omega_j | 1 \leq j \leq J\}$ al conjunto de las clases. En determinadas ocasiones extenderemos Ω con una clase de rechazo ω_0 a la que asignaremos todos aquellos casos para los que no se tiene una certeza aceptable de ser clasificados correctamente en alguna de las clases de Ω. De este modo, denotamos $\Omega^* = \Omega \cup \{\omega_0\}$ al conjunto extendido de clases. Un clasificador o regla de clasificación es una función $d : P \rightarrow \Omega^*$ definida sobre el conjunto de posibles ejemplos P tal que para todo ejemplo X se verifica que $d(X) \in \Omega^*$.

Un modelo de clasificación concreto puede construirse entrevistando a expertos en el tema. De hecho, la construcción de muchos sistemas basados en el conocimiento se basa en la extracción manual del conocimiento de los expertos, a pesar de la dificultad que entraña este proceso. Cuanto mejor es el experto peor suele describir su conocimiento (la paradoja de la Ingeniería del Conocimiento). Además, los expertos en un tema no siempre están de acuer-

do entre sí (la Ley de Hiram: *"Si se consultan suficientes expertos, se puede confirmar cualquier opinión"*).

No obstante, si se dispone de suficiente información registrada (almacenada en una base de datos, por ejemplo), el modelo de clasificación se puede construir generalizando a partir de ejemplos específicos mediante alguna técnica de aprendizaje automático. De hecho, podemos encontrar numerosos ejemplos de algoritmos de aprendizaje automático, como los empleados en la construcción de árboles de decisión (como C4.5 o CART) o los que siguen la metodología STAR de Michalski (INDUCE o AQ, por ejemplo).

Los casos de entrenamiento utilizados en la construcción del modelo de clasificación suelen expresarse en términos de un conjunto finito de propiedades o atributos con valores discretos o numéricos, mientras que las categorías a las que han de asignarse los distintos casos deben establecerse de antemano (al tratarse de aprendizaje supervisado). En general, estas clases serán disjuntas, si bien pueden establecerse jerarquías de clases en las cuales algunas clases son especialización de otras, de modo que las clases no son siempre disjuntas (aunque sí lo suelen ser aquéllas que se encuentran en el mismo nivel de la jerarquía de conceptos empleada para definir las clases del problema). Además, las clases suelen ser discretas, pues si son continuas nos encontramos ante un problema de regresión cuya resolución se puede realizar utilizando técnicas estadísticas. En ocasiones, no obstante, para predecir atributos con valores continuos se definen categorías discretas utilizando términos imprecisos propios del lenguaje natural (esto es, etiquetas lingüísticas que representan conjuntos difusos de valores).

Para que el aprendizaje automático sea correcto, entendiendo éste como un proceso de generalización a partir de ejemplos concretos, hemos de disponer de suficientes casos de entrenamiento (bastantes más que clases diferentes). Si las conclusiones obtenidas no están avaladas por bastantes ejemplos, entonces la aparición de errores en los datos podría conducir al aprendizaje de un modelo erróneo que no resultaría fiable. Por tanto, cuantos más datos obtengamos, más fácilmente podremos diferenciar patrones válidos de patrones debidos a irregularidades o errores.

Por suerte, hoy en día es relativamente fácil recopilar grandes cantidades de

datos relativos al problema que deseemos resolver, pues es sencillo digitalizar información y no resulta excesivamente caro almacenarla. Sin embargo, cuando el tamaño de los conjuntos de datos aumenta considerablemente, muchas de las técnicas tradicionalmente utilizadas en Inteligencia Artificial no resultan adecuadas por ser ineficientes y poco escalables. La necesidad de trabajar eficientemente con grandes conjuntos de datos ha dado lugar al desarrollo de las técnicas de *Data Mining* [17] [75], una rama de las Ciencias de la Computación actualmente en auge [96].

Las técnicas de *Data Mining* se enmarcan dentro del proceso de extracción de conocimiento denominado KDD [59], acrónimo de *Knowledge Discovery in Databases*. Se entiende por KDD la extracción no trivial de información potencialmente útil a partir de un gran volumen de datos en el cual la información está implícita (aunque no se conoce previamente). Su objetivo final es la interpretación de grandes cantidades de datos y el descubrimiento de relaciones o patrones existentes en los datos. Para alcanzar dicho objetivo se emplean algoritmos clásicos de aprendizaje, métodos estadísticos [73] y técnicas de bases de datos [86] [87], lo cual hace del KDD un área de conocimiento eminentemente multidisciplinar.

El proceso de extracción de conocimiento incluye la preparación de los datos y la interpretación de los resultados obtenidos, además de los algoritmos de *Data Mining* propiamente dichos, puesto que de la simple aplicación de técnicas de *Data Mining* sólo se obtienen patrones. Tales patrones no son más que expresiones que describen subconjuntos de los datos de entrada; esto es, son modelos aplicables a los datos de los que se obtuvieron.

Los algoritmos tradicionales de construcción de modelos de clasificación se suelen basar en el descubrimiento de patrones en los datos de entrenamiento y los algoritmos de *Data Mining* proporcionan las técnicas necesarias para poder construir clasificadores de forma eficiente incluso cuando los conjuntos de datos son enormes.

> *Los desarrollos más provechosos han surgido siempre donde se encontraron dos formas de pensar diferentes.*
>
> HEISENBERG

Los árboles de decisión, clasificación o identificación constituyen uno de los modelos más utilizados en aprendizaje supervisado y en aplicaciones de *Data Mining* [68]. Su principal virtud radica en que son modelos de clasificación de fácil comprensión. Además, su dominio de aplicación no está restringido a un ámbito concreto, sino que los árboles de decisión pueden utilizarse en áreas de diversa índole [125], desde aplicaciones de diagnóstico médico hasta sistemas de predicción meteorológica o juegos como el ajedrez.

Los métodos usuales de construcción de árboles de decisión, aun habiéndose utilizado con éxito en incontables ocasiones, carecen de la flexibilidad que ofrecen otras técnicas de inducción de reglas. Éstas, por su parte, aunque pueden llegar a conseguir modelos de clasificación más simples, suelen ser demasiado ineficientes para poder aplicarse con éxito en la resolución de problemas de *Data Mining*.

En esta memoria se pretende diseñar un método que permita construir modelos de clasificación simples, inteligibles y robustos de una forma eficiente y escalable. El objetivo final es lograr un algoritmo eficiente computacionalmente que sea capaz de trabajar con grandes conjuntos de datos, a semejanza de los modernos algoritmos de construcción de árboles de decisión. Estas cualidades se han de conseguir sin olvidar la necesidad de construir clasificadores que destaquen por su simplicidad, tal como los obtenidos mediante algoritmos de inducción de reglas. Por otro lado, el método propuesto ha de ser robusto; es decir, debe ser capaz de funcionar correctamente ante la presencia de ruido en el conjunto de entrenamiento que se utilice para construir el clasificador.

Para cumplir los objetivos identificados en el párrafo anterior, en esta memoria se presenta un método de construcción de modelos de clasificación que combina las mejores cualidades de los árboles de decisión con las de las técnicas de inducción de reglas. El método propuesto da lugar una nueva familia de algoritmos de inducción de árboles de decisión basada en técnicas de extracción de reglas de asociación [19], si bien también puede interpretarse como un algoritmo de inducción de listas de decisión. El uso de técnicas de extracción de reglas de asociación para construir un árbol de decisión da nombre al método propuesto en esta memoria: ART, acrónimo de *Association Rule Trees*.

El modelo de clasificación ART se caracteriza por dotar de mayor flexi-

bilidad a las técnicas tradicionales de construcción de árboles de decisión, al permitir la utilización simultánea de varios atributos para ramificar el árbol de decisión y agrupar en una rama 'else' las ramas del árbol menos interesantes a la hora de clasificar datos, sin descuidar por ello cuestiones relativas a la eficiencia del proceso de aprendizaje. Para lograr dicho propósito se emplean técnicas de extracción de reglas de asociación que nos permiten formular hipótesis complejas de una forma eficiente y métodos de discretización que nos ofrecen la posibilidad de trabajar en dominios continuos.

Como se verá en los capítulos siguientes, la búsqueda de mecanismos complementarios que permitan la aplicación real de ART en distintas situaciones ha dado lugar al desarrollo paralelo de técnicas cuya aplicación no está limitada al modelo de clasificación propuesto, ni siquiera al ámbito del aprendizaje supervisado. En este sentido, además del modelo de clasificación ART (capítulo 3), en esta memoria se presenta un algoritmo eficiente para la extracción de reglas de asociación en bases de datos relacionales (TBAR [19], capítulo 4), un método jerárquico de discretización supervisada (capítulo 5) y una arquitectura distribuida basada en componentes apta para cualquier aplicación de cálculo intensivo (capítulo 6).

Contenido de esta memoria

En el siguiente capítulo se presentan algunos conceptos básicos relacionados con las técnicas en que se basa el modelo de clasificación ART. En concreto, se estudia la construcción de árboles de decisión, la inducción de listas de decisión y la extracción de reglas de asociación, así como el uso de estas últimas en la resolución de problemas de clasificación.

Los algoritmos de construcción de árboles de decisión, estudiados en la sección 2.1, suelen construir de forma descendente los árboles de decisión, comenzando en la raíz del árbol. Por este motivo se suele hacer referencia a este tipo de algoritmos como pertenecientes a la familia TDIDT [*Top-Down Induction of Decision Trees*]. En este sentido, ART es un algoritmo TDIDT más, si bien el criterio de preferencia utilizado para ramificar el árbol de decisión y la topología del mismo difieren notablemente de las propuestas anteriores. Es precisamente la topología del árbol construido por ART la que permite interpretar el modelo de clasificación construido por ART como una lista de decisión. Las listas de decisión, como caso particular de los algoritmos de inducción de reglas, se analizan en la sección 2.2.

El capítulo 2 se cierra con la sección 2.3, en la que se estudian las técnicas de extracción de reglas de asociación y se comenta su posible uso en la construcción de modelos de clasificación.

Tras analizar los tres pilares en los que se asienta el modelo propuesto en esta memoria (árboles de clasificación, listas de decisión y reglas de asociación), el capítulo 3 se dedica por completo a la presentación del modelo de clasificación ART.

En la sección 3.1 se describe el algoritmo ART para la construcción de árboles de decisión. Acto seguido, en el apartado 3.2, aparece un ejemplo detallado que ayuda a comprender el funcionamiento del proceso de inducción llevado a cabo por ART. A continuación, la utilización de ART en la resolución de un problema real es objeto de la sección 3.3. Tras estos ejemplos concretos, que ayudan a entender mejor el modelo de clasificación ART, en la sección 3.4 se analiza el uso de los clasificadores ART (esto es, los modelos de clasificación obtenidos como resultado de aplicar el algoritmo ART).

El método ART obtiene modelos de clasificación excelentes a la vez que es capaz de trabajar adecuadamente con los enormes conjuntos de datos que suelen utilizarse para resolver problemas de extracción de conocimiento en bases de datos y *Data Mining*. Esta cualidad se debe a las buenas propiedades de escalabilidad de los algoritmos de extracción de reglas de asociación que utiliza internamente para ramificar el árbol de decisión. Éstas y otras propiedades del modelo de clasificación ART se comentan en la sección 3.5.

El tercer capítulo de esta memoria se cierra en el apartado 3.6 con la presentación de los resultados experimentales que se han obtenido con el modelo de clasificación propuesto. Como caso particular, se han obtenido empíricamente resultados interesantes al utilizar ART para clasificar uniones de genes en secuencias de ADN, problema en el que ART descubre y aprovecha las relaciones existentes entre los nucleótidos de una secuencia de ADN (véase la figura 3.1 de la página 63, que muestra el clasificador ART obtenido para este problema, cuya descripción aparece en el apartado 3.3 de esta memoria).

Si bien ART no siempre mejora los porcentajes de clasificación obtenidos al utilizar otros clasificadores existentes, como C4.5 [131] o RIPPER [38], los clasificadores ART suelen comportarse bien en términos de porcentaje de clasificación, simplicidad y robustez ante la presencia de ruido en los datos de entrenamiento. Además, ART dota de mayor flexibilidad al proceso tradicional de construcción de árboles de decisión y ofrece una alternativa escalable a las técnicas existentes de inducción de listas de decisión.

Como ya se ha comentado, el modelo de clasificación propuesto en esta memoria emplea internamente un eficiente algoritmo de extracción de reglas de asociación. Dicho algoritmo, denominado TBAR [19], es el núcleo en torno al cual gira el capítulo 4 de esta memoria.

Este algoritmo de extracción de reglas de asociación, como parte de ART, ofrece un mecanismo simple y efectivo para tratar una amplia variedad de situaciones sin necesidad de recurrir a otras técnicas más específicas, complejas y artificiales. Además, su eficiencia y escalabilidad permiten que ART sea perfectamente capaz de trabajar con los enormes conjuntos de datos comunes en problemas de *Data Mining*.

El algoritmo TBAR, como método general de extracción eficiente de reglas

de asociación en bases de datos relacionales, es objeto de estudio en la sección 4.2, mientras que su uso en ART se analiza en la sección 4.3.

El capítulo dedicado a la extracción de reglas de asociación también incluye un apartado, 4.4, en el que se describen distintas medidas que se pueden emplear para evaluar las reglas obtenidas por algoritmos como TBAR.

El algoritmo TBAR, en concreto, está diseñado para mejorar el rendimiento de otros métodos alternativos de extracción de reglas de asociación. De hecho, TBAR permite reducir al máximo los recursos computacionales necesarios para llevar a cabo este proceso, tanto el tiempo de ejecución requerido como el espacio de almacenamiento consumido.

Las reglas que se utilizan para construir el modelo de clasificación ART se obtienen a partir de conjuntos de datos que usualmente contendrán atributos de tipo numérico. Para no limitar innecesariamente la aplicabilidad de técnicas de aprendizaje como ART, se necesitan mecanismos que permitan construir modelos de clasificación con atributos continuos. El capítulo 5 de esta memoria se centra precisamente en el estudio de las técnicas que nos permiten construir árboles de decisión con atributos numéricos.

En la sección 5.1 se analizan las técnicas de discretización existentes, las cuales permiten que cualquier técnica de aprendizaje pueda trabajar con atributos continuos tratándolos como si fueran categóricos. Dado que muchos de los métodos de discretización existentes no aprovechan la información de que disponen, en la sección 5.2, se propone un nuevo método de discretización, el discretizador contextual, que resulta especialmente adecuado cuando se utiliza durante la construcción de modelos de clasificación.

La sección 5.3 está dedicada a la construcción de árboles de decisión con atributos de tipo numérico y en ella se describe cómo puede utilizarse el discretizador contextual de la sección 5.2 para dotar de mayor flexibilidad al proceso de construcción del árbol de decisión.

Los resultados experimentales que se han obtenido al utilizar distintos métodos de discretización se discuten en la sección 5.4, donde se analiza el uso de estas técnicas en la construcción de árboles de decisión TDIDT (apartado 5.4.1) y su aplicación en el modelo de clasificación ART (apartado 5.4.2).

Las principales aportaciones del capítulo 5 son, en primer lugar, la apli-

cación de un método de discretización jerárquico al proceso de construcción de árboles de decisión con atributos continuos y, en segundo lugar, la presentación de un método de discretización alternativo que procura hacer uso de la información disponible para conseguir una partición óptima del dominio de un atributo continuo.

Tras haber estudiado el modelo de clasificación ART y las dos técnicas que lo hacen útil en problemas reales (esto es, la extracción eficiente de reglas de asociación y la discretización de valores continuos), el capítulo 6 de esta memoria se centra en el estudio de la infraestructura computacional que hace factible la aplicación de técnicas de *Data Mining* a gran escala (como puede ser la construcción de clasificadores ART). En concreto, en este capítulo se propone la implementación de un sistema distribuido basado en componentes.

El modelo conceptual de un sistema general de *Data Mining* se presenta en la sección 6.1. Tal sistema sería conveniente que fuese distribuido dadas las necesidades computacionales de las técnicas que trabajan con enormes cantidades de datos. En el apartado 6.2 se propone un sistema descentralizado que pueda hacer frente a dichas necesidades.

La infraestructura planteada en el capítulo 6 debe proporcionar servicios de forma dinámica, por lo cual es recomendable implementarla como un sistema basado en componentes, fácilmente adaptable y reconfigurable, con la arquitectura propuesta en la sección 6.3.

El capítulo dedicado a cuestiones de infraestructura incluye, además, una descripción en el apartado 6.4 de los criterios que se habrían de seguir al diseñar un sistema como el descrito y de cómo se pueden implementar los subsistemas de los que se compone utilizando tecnologías existentes en la actualidad.

La infraestructura propuesta resulta de interés, no sólo para resolver problemas de *Data Mining*, sino para cualquier tipo de aplicación de cálculo intensivo para las que usualmente se recurre al uso de costosos supercomputadores.

Esta memoria se cierra con el capítulo 7, en el cual se exponen algunas de las conclusiones a las que se ha llegado durante el desarrollo de este trabajo, así como algunas sugerencias encaminadas a la realización de futuros trabajos.

En cuanto a los resultados que aparecen reflejados en esta memoria, además del modelo de clasificación ART analizado en el capítulo 3, resulta digno de

mención el hecho de que se hayan desarrollado algunas ideas y técnicas cuyo ámbito de aplicación va más allá de su uso específico en la construcción de clasificadores ART. Entre ellas destacan el algoritmo TBAR de extracción de reglas de asociación en bases de datos relacionales, el método de discretización contextual y la arquitectura distribuida basada en componentes propuesta para la resolución de problemas de cálculo intensivo. Descripciones detalladas de dichas propuestas pueden encontrarse en los capítulos 4, 5 y 6 de la presente memoria, respectivamente.