Capítulo 7

Conclusiones

Somos la suma de nuestras decisiones

WOODY ALLEN
Delitos y Faltas (1989)

En este trabajo se ha presentado una nueva estrategia para construir árboles de decisión que ha conseguido resultados prometedores experimentalmente: el modelo de clasificación ART. En realidad, este modelo de clasificación, que se presenta en el capítulo 3 de la presente memoria, se puede considerar un híbrido entre los algoritmos de construcción de árboles de decisión y los algoritmos de inducción de listas de decisión, técnicas de aprendizaje supervisado que se describen en las secciones 2.1 y 2.2 respectivamente.

Como algoritmo de construcción de árboles de decisión, ART se caracteriza por construir árboles de decisión n-arios y ser capaz de utilizar simultáneamente varios atributos para ramificar el árbol de decisión, lo que le permite construir modelos de clasificación más compactos.

En cuanto a la interpretación del método ART como algoritmo de inducción de listas de decisión, cabe destacar que ART mejora la eficiencia de propuestas anteriores al extraer en paralelo conjuntos de reglas que sirven para ramificar cada nivel del árbol de decisión, mientras que las técnicas habituales de inducción de listas de decisión se limitan a ir descubriendo de una en una

las reglas que forman parte del modelo de clasificación construido por ellas.

En definitiva, como árbol de decisión, ART ofrece un modelo más flexible que CART o C4.5 al construir árboles n-arios con ramas 'else', mientras que como algoritmo de inducción de listas de decisión obtiene un modelo de menor profundidad, con lo que resulta más fácil de interpretar, y además lo hace de una forma más eficiente gracias a las técnicas empleadas en su implementación.

La implementación de ART se basa en la utilización de una técnica eficiente de extracción de reglas de asociación. Esta técnica, habitual en aplicaciones de *Data Mining*, es la que da nombre al modelo de clasificación propuesto en esta memoria, ya que ART es acrónimo de *Association Rule Tree*.

Gracias al uso de eficientes algoritmos de extracción de reglas de asociación, la complejidad del proceso de construcción del clasificador ART es comparable a la de los populares algoritmos TDIDT de construcción de árboles de decisión y puede llegar a ser varios órdenes de magnitud más eficiente que algoritmos de inducción de reglas como CN2 o RIPPER, especialmente cuando aumenta el tamaño de los conjuntos de datos (véanse, por ejemplo, las figuras de la página 103).

Al estar basado en algoritmos de extracción de reglas de asociación, ART no sólo es eficiente, sino que también es escalable. Ésta es una característica esencial en la resolución de problemas de *Data Mining*, pues permite la utilización del método propuesto en esta memoria para extraer información de enormes conjuntos de datos.

El algoritmo de extracción de reglas de asociación utilizado por ART también 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. En concreto, el proceso de extracción de reglas de asociación es responsable de que ART se comporte correctamente ante la presencia de información con ruido o la existencia de claves primarias en el conjunto de entrenamiento, ya que ART se basa en la extracción de itemsets frecuentes para construir el modelo de clasificación.

Por otro lado, la topología particular del árbol construido por ART facilita tratar de una forma elegante la presencia de valores desconocidos para los

atributos involucrados en el test de un nodo interno del árbol: cuando se le presenta un caso de prueba con valores desconocidos y no es aplicable el test que dio lugar a la ramificación del árbol, ART simplemente envía el caso de prueba a la rama 'else' del nodo.

Debido en parte a la topología del árbol, ART es capaz de construir clasificadores que destacan por su simplicidad e inteligibilidad, además de por su robustez ante la presencia de ruido. ART se caracteriza por construir modelos de clasificación de menor complejidad que los obtenidos mediante la utilización de algoritmos TDIDT y de complejidad similar a las listas de decisión que se obtienen empleando algoritmos más ineficientes. Es más, la simplicidad de los modelos de clasificación ART se consigue sin sacrificar la precisión del clasificador, lo que hace de ART una alternativa interesante en situaciones en las que el tamaño de los árboles de decisión TDIDT los hace prácticamente inmanejables.

Como caso particular, ART logra resultados interesantes cuando se utiliza para clasificar uniones de genes en secuencias de ADN, tal como se muestra en la sección 3.3. En este problema, ART descubre y aprovecha las simetrías existentes en las secuencias de nucleótidos alrededor de una unión para construir un modelo de clasificación mucho más sencillo que el que se obtiene utilizando otros métodos de construcción de árboles de decisión.

En cuanto a la utilización de ART por parte de usuarios no expertos, es destacable el hecho de que ART requiere un conjunto limitado de parámetros que usualmente no hay que ajustar. Básicamente, los parámetros empleados por ART son los habituales en cualquier proceso de extracción de reglas de asociación y sirven para acotar el espacio de búsqueda explorado al construir el clasificador.

Uno de los parámetros habituales en el proceso de extracción de reglas de asociación (y, por tanto, en ART) es el umbral mínimo de confianza que se le exige a las reglas de asociación para ser consideradas como hipótesis candidatas en la construcción del árbol de decisión. Este parámetro puede utilizarse, por ejemplo, para establecer una restricción a priori sobre la precisión de las reglas seleccionadas para formar parte del clasificador. Esta posibilidad es muy interesante en la resolución de algunos problemas y necesaria cuando no

se permiten errores, como en el caso de la identificación de setas comestibles comentado en la sección 3.1.3 de esta memoria.

En la misma sección del presente trabajo, también se ha expuesto el uso de técnicas heurísticas que permiten seleccionar automáticamente el mejor conjunto de reglas descubiertas, evitando de este modo que el usuario tenga que establecer manualmente los parámetros empleados en la construcción del clasificador ART. En particular, se propone la utilización de un margen de tolerancia en la selección de reglas candidatas a formar parte del árbol ART. Esta heurística concreta consigue resultados excepcionales sin que el usuario tenga que estimar los valores adecuados para los parámetros utilizados en la construcción de clasificadores ART.

En la sección 4.4 de esta memoria, se presentan alternativas al uso de la confianza como criterio de estimación de la calidad de las reglas obtenidas. Si bien las medidas analizadas no consiguen resultados netamente superiores a los obtenidos al utilizar la confianza, medidas como los factores de certeza o el interés de las reglas pueden resultar útiles en situaciones particulares dependiendo de la semántica del problema que se desee resolver.

De hecho, del estudio de medidas alternativas para evaluar las reglas extraídas surgió la idea de imponer restricciones adicionales a las reglas empleadas para construir el clasificador ART. En el apartado 4.4.3.12 se describe un criterio que permite mejorar el porcentaje de clasificación obtenido por ART a cambio de un pequeño aumento en la complejidad del clasificador construido.

En lo que respecta al uso en la práctica del modelo de clasificación ART, también hay que mencionar que se ha comprobado experimentalmente su buen funcionamiento cuando se emplean técnicas de discretización. Estas técnicas permiten construir clasificadores ART con atributos continuos, algo esencial si deseamos aplicar el modelo de clasificación ART en la resolución de problemas reales. Al realizar los experimentos que aparecen recogidos en la sección 5.4.2, se llegó a la conclusión de que ART funciona mejor cuando se emplean técnicas de discretización de forma global (antes de comenzar la construcción del clasificador), pues el uso de técnicas de discretización local (en cada nodo del árbol) se traduce generalmente en clasificadores de construcción más costosa computacionalmente, más complejos en cuanto a su tamaño y algo menos

precisos que los obtenidos realizando una discretización global de los atributos continuos del conjunto de entrenamiento.

Aparte del modelo de clasificación ART, cuyas mejores cualidades se han comentado en los párrafos anteriores, durante el desarrollo del trabajo descrito en esta memoria se han obtenido resultados adicionales que han dado lugar a una serie de 'subproductos' con entidad propia, cuya utilidad va más allá de su uso en ART. Algunos de los más relevantes se describen a continuación.

El algoritmo TBAR de extracción de reglas de asociación

El algoritmo TBAR [19], presentado en la sección 4.2 de esta memoria, es un algoritmo eficiente de extracción de reglas de asociación que resulta especialmente adecuado para su utilización en conjuntos de datos densos, como los que se pueden encontrar en bases de datos relacionales y se emplean habitualmente para resolver problemas de clasificación. El algoritmo TBAR, como técnica general de extracción de reglas de asociación, mejora el rendimiento de otros algoritmos ampliamente utilizados, como es el caso de Apriori [7].

El discretizador contextual

En la sección 5.2 se presentó este método de discretización, jerárquico y supervisado si nos atenemos a las categorías descritas en el capítulo 5. Este método, que obtiene buenos resultados cuando se utiliza para construir clasificadores TDIDT y ART, utiliza la estrategia tradicional de los métodos de agrupamiento clásicos para discretizar los valores de un atributo continuo. El discretizador contextual emplea medidas de similitud entre intervalos adyacentes (cualquiera de las que aparecen en el anexo 5.5) en vez de utilizar medidas de pureza como suelen hacer los demás métodos existentes de discretización supervisada.

Reglas de división alternativas para la construcción de árboles de decisión con algoritmos TDIDT

Las medidas de pureza son las que se suelen utilizar como reglas de división para decidir cómo se ramifica un árbol de decisión, tal como se puede leer

en el apartado 2.1.1. En dicho apartado se mencionaron dos medidas, MAX-DIF y el Índice Generalizado de Gini, que son de formulación más sencilla que los demás criterios existente y obtienen resultados comparables con los conseguidos utilizando reglas de división más complejas [18].

Árboles n-arios arbitrarios utilizando técnicas discretización jerárquica en algoritmos TDIDT

También se ha descrito en esta memoria el empleo de técnicas de discretización jerárquica para construir árboles n-arios arbitrarios con atributos numéricos, árboles en los cuales no se restringe el factor de ramificación del árbol final (sección 5.3.2).

Una arquitectura distribuida de cómputo basada en componentes

Finalmente, en el capítulo 6 se plantea una arquitectura adecuada para la resolución de problemas de cómputo intensivo, como puede ser la construcción de clasificadores ART en aplicaciones de *Data Mining*.

La arquitectura propuesta, distribuida al estilo de los sistemas P2P actuales, está basada en componentes para facilitar el desarrollo de aplicaciones que hagan uso de la infraestructura que ofrece.

Además, esta arquitectura general incluye dos subsistemas cuyo ámbito de aplicación va más allá de los límites del sistema planteado:

- El modelo propuesto para los conjuntos de datos con los que trabaja el sistema (sección 6.4.2) se puede utilizar en cualquier sistema que tenga que manipular conjuntos de datos o acceder a distintas fuentes de datos de una manera uniforme aunque éstas sean heterogéneas.

- El almacén de información propuesto al describir el servicio de persistencia del sistema basado en componentes (sección 6.4.3) también puede ser útil en un amplio rango de aplicaciones, al poder almacenar objetos de cualquier tipo sin tener que modificar la estructura de la base de datos subyacente (funcionando de una forma similar al catálogo de una base de datos relacional).

Trabajo futuro

Una vez que se han descrito los resultados más relevantes que se han obtenido durante la realización de este trabajo, se sugiere una serie de puntos de partida para líneas de investigación futuras:

- Resulta de especial interés estudiar la posibilidad de construir un modelo de clasificación híbrido TDIDT-ART que seleccione la estrategia de ART o de los algoritmos TDIDT en función de la situación que se le presente en cada nodo durante la construcción de un árbol de decisión.

- También es interesante la incorporación de técnicas difusas en el modelo de clasificación ART, para lo cual se necesitan algoritmos de extracción de reglas de asociación difusas.

- Igualmente, puede resultar de provecho investigar formas alternativas de manejar atributos continuos, como puede ser la utilización de técnicas de extracción de reglas de asociación cuantitativas [146] [114] [1] [12].

- Del mismo modo, la extensión de ART para tratar problemas de regresión (esto es, cuando los atributos continuos aparecen en el consecuente de las reglas) es otra línea de trabajo futuro.

- El proceso de extracción de reglas candidatas es otra faceta de ART que merece un estudio adicional para encontrar técnicas que obtengan mejores reglas de una forma más eficiente.

- La introducción de pesos en las reglas puede ser otra estrategia que permita obtener clasificadores ART más precisos.

- El estudio de medidas alternativas de evaluación de las reglas obtenidas en siempre una línea de trabajo abierta que puede conducir a la consecución de buenos resultados.

- También puede ser interesante el análisis de criterios de preferencia alternativos que nos permitan seleccionar las mejores reglas del conjunto de reglas disponible para construir cada nivel del árbol ART.

- Un estudio más en profundidad acerca de la posible utilización de ART como técnica de aprendizaje incremental también es deseable, ya que la existencia de métodos eficientes que permitan actualizar un clasificador dado un conjunto de actualizaciones de una base de datos es muy importante en aplicaciones de *Data Mining*.

- De forma complementaria, puede ser beneficioso el empleo de técnicas de post-procesamiento que permitan mejorar el rendimiento del clasificador ART una vez construido.