



**UNIVERSIDAD DE GRANADA
E.T.S. INGENIERÍA INFORMÁTICA**

**Departamento de
Ciencias de la Computación
e Inteligencia Artificial**

TESIS DOCTORAL

ART

**Un método alternativo
para la construcción de árboles de decisión**

Fernando Berzal Galiano

Granada, junio de 2002



ART
Un método alternativo
para la construcción de árboles de decisión

memoria que presenta

Fernando Berzal Galiano

para optar al grado de

Doctor en Informática

Junio de 2002

DIRECTOR

Juan Carlos Cubero Talavera

DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

E INTELIGENCIA ARTIFICIAL

E.T.S. INGENIERÍA INFORMÁTICA

UNIVERSIDAD DE GRANADA

La memoria titulada “ART: Un método alternativo para la construcción de árboles de decisión”, que presenta D. Fernando Berzal Galiano para optar al grado de Doctor, ha sido realizada en el Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada bajo la dirección del Doctor Juan Carlos Cubero Talavera.

Granada, junio de 2002.

El Doctorando

El Director

Fdo. Fernando Berzal

Fdo. Juan Carlos Cubero

Agradecimientos

En primer lugar, he de reconocer el esfuerzo, tesón y dedicación de una persona muy especial para mí, mi madre Adelaida, que siempre me ha apoyado en mis decisiones y ha estado ahí en los buenos momentos y en los no tan buenos.

En segundo lugar, pero no por ello de menor importancia, tengo que agradecerle a mi director, Juan Carlos, el interés que ha mostrado por mí desde que hace ya algunos años fui a pedirle una carta de recomendación, tras lo cual acabé siendo “su” becario e hice con él mi Proyecto de Fin de Carrera. Al fin y al cabo, la mera existencia de esta memoria se debe a su persistencia (y también a su paciencia). Durante estos años me ha demostrado su calidad como tutor y, sobre todo, su valía como persona. Espero que en el futuro, pase lo que pase, tenga en él a un gran amigo.

Así mismo, les debo mucho a las personas con las cuales he desarrollado mi trabajo en el seno del Departamento de Ciencias de la Computación e Inteligencia Artificial de la Universidad de Granada, en particular a mis compañeros de mudanzas de Mecenaz y a los integrantes del grupo de investigación IDBIS, desde su diligente y solícita directora hasta el artista del grupo. Mención especial merece mi compañero de despacho, amigo y socio. En cierto sentido, Nicolás se ha convertido en algo así como un hermano mayor para mí (entre otras cosas, por su perspectiva desde el año de ventaja que me lleva).

Por otro lado, tampoco puedo olvidar a los muchos profesores que han ido guiando mi desarrollo académico y personal. Aunque de pequeño quería ser profesor de Geografía e Historia como mi abuelo, del que heredé mi fascinación por los libros, mi inclinación por las Ciencias no tardó demasiado en aparecer. De hecho, mis primeros devaneos con las Matemáticas provienen de mi etapa en EGB con un profesor excepcional, Fernando Barranco, Sch.P., alguien inolvidable para muchos estudiantes que pasamos por los Escolapios de Granada. Allí conocí a profesores inigualables que son, además, bellísimas personas. En concreto, me estoy refiriendo a Mari Carmen López del Amo y a Fernando Martín, dos profesores a los que siempre recordaré con cariño.

Va por todos ellos...

PD: Aparte de las personas mencionadas y de aquellas a las que haya podido omitir, he de confesar la colaboración de ELVEX, que ha cumplido sobradamente a pesar de sus frecuentes idas y venidas, y también del viejo SHERLOCK, el cual ha realizado su trabajo a duras penas, si bien es cierto que nunca me ha fallado. ¡Ah! Casi me olvido de mi obsoleto CPC, al cual le debo mi pasión por la Informática ;-)

Índice general

1. Introducción	1
2. Propedéutica	13
2.1. Árboles de decisión	15
2.1.1. Reglas de división	18
2.1.1.1. Ganancia de información: Entropía	19
2.1.1.2. El criterio de proporción de ganancia	21
2.1.1.3. El índice de diversidad de Gini	22
2.1.1.4. Otros criterios	23
2.1.2. Reglas de parada	26
2.1.3. Reglas de poda	27
2.1.3.1. Poda por coste-complejidad	28
2.1.3.2. Poda pesimista	29
2.1.4. Algoritmos TDIDT	30
2.1.5. Paso de árboles a reglas	35
2.2. Inducción de reglas y listas de decisión	36
2.2.1. Metodología STAR	37
2.2.2. Listas de decisión	41
2.2.3. Algoritmos genéticos	45
2.3. Reglas de asociación	48
2.3.1. Algoritmos de extracción de reglas de asociación	50
2.3.2. Construcción de clasificadores con reglas de asociación	58
3. El modelo de clasificación ART	61
3.1. Construcción del clasificador	64
3.1.1. Selección de reglas: Criterio de preferencia	67

3.1.2.	Topología del árbol: Ramas 'else'	70
3.1.3.	Extracción de reglas: Hipótesis candidatas	74
3.1.3.1.	El umbral de soporte mínimo: <i>MinSupp</i>	75
3.1.3.2.	El umbral de confianza mínima: <i>MinConf</i>	76
3.2.	Un ejemplo detallado	78
3.3.	Un caso real: SPLICE	83
3.4.	Notas acerca del clasificador ART	85
3.4.1.	Clasificación con ART	85
3.4.2.	Manejo de valores nulos	86
3.4.3.	Conversión del árbol en reglas	87
3.5.	Propiedades del clasificador ART	89
3.5.1.	Estrategia de búsqueda	89
3.5.2.	Robustez (ruido y claves primarias)	90
3.5.3.	Complejidad del árbol	91
3.6.	Resultados experimentales	91
3.6.1.	Precisión	92
3.6.2.	Eficiencia	98
3.6.3.	Complejidad	102
3.6.4.	El umbral de confianza	105
3.6.5.	El umbral de soporte	106
3.6.6.	Otros experimentos	111
3.6.7.	Comentarios finales	112
4.	Construcción de hipótesis candidatas	113
4.1.	Extracción de reglas de asociación	115
4.1.1.	El algoritmo Apriori	115
4.1.2.	El algoritmo DHP	117
4.2.	El algoritmo \overline{T} (TBAR)	118
4.2.1.	Visión general de TBAR	119
4.2.1.1.	Obtención de los itemsets relevantes	120
4.2.1.2.	Generación de las reglas de asociación	121
4.2.2.	El árbol de itemsets	122
4.2.2.1.	Inicialización del árbol de itemsets	125
4.2.2.2.	Obtención de los itemsets relevantes	126
4.2.2.3.	Generación de candidatos	127
4.2.2.4.	Derivación de reglas de asociación	128

4.2.3.	Resultados experimentales	132
4.2.4.	Observaciones finales sobre TBAR	139
4.3.	\bar{T} en ART: Reglas de asociación con restricciones	141
4.3.1.	Extracción de itemsets	141
4.3.2.	Generación de reglas	143
4.4.	Evaluación de las reglas obtenidas	144
4.4.1.	Propiedades deseables de las reglas	145
4.4.2.	Medidas de relevancia de un itemset	146
4.4.2.1.	Soprote	146
4.4.2.2.	Fuerza colectiva	147
4.4.3.	Medidas de cumplimiento de una regla	148
4.4.3.1.	Confianza	148
4.4.3.2.	Confianza causal	149
4.4.3.3.	Soprote causal	149
4.4.3.4.	Confirmación	150
4.4.3.5.	Convicción	152
4.4.3.6.	Interés	153
4.4.3.7.	Dependencia	154
4.4.3.8.	Dependencia causal	154
4.4.3.9.	Medida de Bhandari	155
4.4.3.10.	Divergencia Hellinger	155
4.4.3.11.	Factores de certeza	155
4.4.3.12.	Cuestión de utilidad	161
4.4.4.	Resultados experimentales	162
5.	Manejo de atributos continuos	169
5.1.	La discretización de espacios continuos	170
5.1.1.	El caso general: Métodos de agrupamiento	170
5.1.2.	El caso unidimensional: Métodos de discretización	173
5.1.2.1.	Clasificación	173
5.1.2.2.	Algoritmos de discretización	175
5.2.	Discretización contextual: Un enfoque alternativo	176
5.2.1.	La discretización contextual como método de discretización supervisada	176
5.2.2.	La discretización contextual como método de discretización jerárquica	178

5.2.2.1.	Discretización contextual aglomerativa . . .	178
5.2.2.2.	Discretización contextual divisiva	179
5.2.2.3.	Eficiencia de la discretización contextual . .	181
5.2.3.	Uso de la discretización contextual como método de discretización local	182
5.2.4.	Un pequeño ejemplo	183
5.3.	Atributos continuos en árboles de decisión	188
5.3.1.	Árboles binarios vs. árboles n-arios	188
5.3.1.1.	Árboles binarios con atributos continuos . .	189
5.3.1.2.	Árboles n-arios con atributos continuos . . .	191
5.3.2.	Discretización local jerárquica en árboles n-arios . . .	192
5.3.2.1.	Versión aglomerativa	193
5.3.2.2.	Variante divisiva	195
5.3.2.3.	Eficiencia	196
5.4.	Resultados experimentales	197
5.4.1.	Discretización en algoritmos TDIDT	199
5.4.1.1.	Discretización local	202
5.4.1.2.	Discretización global	207
5.4.2.	ART con discretización	211
5.4.2.1.	Precisión	211
5.4.2.2.	Complejidad	211
5.4.3.	Observaciones finales	219
5.5.	Anexo: Medidas de similitud	222
5.5.1.	Modelos basados en medidas de distancia	223
5.5.2.	Modelos basados en medidas de correlación	224
5.5.3.	Modelos basados en Teoría de Conjuntos	225
5.5.4.	Resultados experimentales	229
6.	Cuestión de infraestructura	235
6.1.	Modelo conceptual del sistema	238
6.2.	Sistemas distribuidos	242
6.2.1.	Evolución y tendencias	243
6.2.1.1.	Sistemas P2P	243
6.2.1.2.	La taxonomía IFCS	248
6.2.2.	Requisitos del sistema	249
6.2.2.1.	Comunicación entre nodos de procesamiento	249

6.2.2.2.	Acceso a los datos	252
6.2.2.3.	Dinámica del sistema	254
6.2.2.4.	Seguridad y fiabilidad	255
6.2.2.5.	Planificación y asignación de recursos	256
6.3.	Sistemas basados en componentes	257
6.3.1.	Patrón de diseño	258
6.3.2.	El kernel del sistema	262
6.4.	Diseño e implementación	264
6.4.1.	Principios de diseño	266
6.4.1.1.	Transparencia	266
6.4.1.2.	Usabilidad	267
6.4.1.3.	Patrones de diseño	267
6.4.2.	Modelización de conjuntos de datos	268
6.4.3.	Servicio de persistencia	273
6.4.4.	Implementación del sistema en Java	276
6.4.5.	Despliegue del sistema	280
6.5.	Una mirada hacia el futuro	280
7.	Conclusiones	283