Índice

Resumen 15

Motivación 15

Desarrollos y aportes 16

Publicaciones derivadas de esta tesis doctoral 19

Capítulo 1. Introducción a la Minería de Datos 21
1. Minería de datos 21
   1.1. Tipos de datos 24
   1.2. Tipos de modelos 26
2. Extracción de conocimiento 26
   2.1. Fase de integración y recopilación 28
   2.2. Fase de selección, limpieza y transformación 29
      2.2.1. Limpieza y transformación 30
         2.2.1.1. Discretización 32
         2.2.1.2. Numerización 32
         2.2.1.3. Normalización de rango: escalado y centrado 33
      2.2.2. Exploración y selección 33
   2.3. Fase de minería de datos 34
      2.3.1. Tareas predictivas 36
      2.3.2. Tareas descriptivas 37
      2.3.3. Técnicas 38
      2.3.4. Aprendizaje inductive 40
      2.3.5. Grandes bases de datos 40
   2.4. Fase de evaluación e interpretación 41
      2.4.1. Técnicas de evaluación 42
         2.4.1.1. Validación simple 42
         2.4.1.2. Validación cruzada con k pliegues 42
         2.4.1.3. Bootstrapping 42
      2.4.2. Medidas de evaluación de modelos 43
      2.4.3. Interpretación y contextualización 44
   2.5. Fase de difusión, uso y monitorización 44
3. Arboles de decisión 46

3.1. Particiones 47
3.2. Criterio de selección de particiones 49
3.3. Poda y reestructuración 50
3.4. Extracción de reglas 51
4. Algoritmos evolutivos 52
5. Minado de datos incremental 55
5.1. Adaptabilidad del modelo 56
6. Toma de decisiones 57
7. Hiper-rectángulos 58
7.1. El uso de los hiper-rectángulos en minería de datos 59

Capítulo 2. Clasificación utilizando hiper-rectángulos. Armado del modelo de datos y obtención de reglas de clasificación 61
1. Hiper-rectángulos 62
1.1. Creación de hiper-rectángulos a partir de una base de datos 64
2. Superposiciones 66
2.1. Tipos de superposiciones 68
2.1.1. Superposición sin datos involucrados 68
2.1.2. Superposición con datos de una clase 69
2.1.3. Superposición con datos de ambas clases 71
2.2. Eliminación de superposiciones 73
2.2.1. Sin datos involucrados 75
2.2.2. Con datos de una clase en la superposición 76
2.2.3. Con datos de ambas clases 77
3. Índices 78
3.1. Índices de superposición 80
3.1.1. Z1i – Proporción del ancho de la intersección de área respecto al ancho del hiper-rectángulo 81
3.1.2. Z2i – Proporción del ancho del intervalo de la intersección de datos con respecto al ancho del intervalo del subconjunto de datos participante 81
3.1.3. Z3i – Proporción del ancho del intervalo del subconjunto de datos intersectados en relación al ancho del intervalo del subconjunto de datos participante 83
3.1.4. Z4i – Proporción del ancho del intervalo del subconjunto de datos participantes en relación al ancho de la superposición de área 84
3.1.5. Z5i – Desplazamiento del intervalo del subconjunto de datos intersectados de un hiper-rectángulo en relación al mínimo del intervalo de subconjunto de datos participantes del otro hiper-rectángulo. 85

3.1.6. Z6i – Desplazamiento del intervalo del subconjunto de datos intersectados de un hiper-rectángulo en relación al máximo del intervalo de subconjunto de datos participantes del otro hiper-rectángulo. 88
3.2. Índice de separabilidad Ω 89
3.2.1. Ponderando por la cantidad de datos participantes 91
3.2.1.1. Z1i 92
3.2.1.2. Z2i 92
3.2.1.3. Z3i 92
3.2.1.4. Z4i 92
3.2.1.5. Z5i 94
3.2.1.6. Z6i 94
3.2.1.7. Re-definición del cálculo de Ωi ponderado por los pesos V 94
3.2.2. Ponderando los índices por otros criterios 95
3.3. Una estrategia de clasificación flexible 95
4. CLUHR 96
4.1. Inicialización del algoritmo 99
4.1.1. Detectar superposiciones iniciales 100
4.2. Eliminar todas las superposiciones 100
4.2.1. Calcular los índices Ω 100
4.2.2. Realizar el ajuste 101
4.2.2.1. Método alternativo para la división de hiper-rectángulos cuando hay datos de ambas clases en la superposición 101
4.2.3. Actualizar los hiper-rectángulos representativos mínimos 102
4.2.4. Detectar las nuevas superposiciones 103
4.3. Finalizar con el armado del modelo de datos 105
4.4. Estructura del modelo de datos 106
4.5. Datos faltantes 106
4.6. Una metodología determinista 107
4.7. Limitaciones de CLUHR 108
5. Extracción de las reglas 109
5.1. Método greedy 111
6. Uso del modelo. Predicción 112
7. Intervención del experto 115

Capítulo 3. Adaptabilidad y actualización del modelo de datos 117
1. Adaptabilidad del modelo 118
1.1. Precondiciones 119
2. Actualización en línea 119

2.1. Agregando nuevos datos 120
2.1.1. El nuevo dato está incluido en un único hiper-rectángulo 121
2.1.2. El nuevo dato está incluido en una superposición entre dos hiper-rectángulos 122
2.1.3. El nuevo dato no está incluido en ningún hiper-rectángulo 124
2.2. Eliminando datos existentes 127
2.2.1. El dato está incluido en un hiper-rectángulo representante de otra clase 128
2.2.2. El dato está incluido en un hiper-rectángulo representante de su misma clase 128
2.3. Modificación de la clase de los datos 130
2.3.1. El dato está incluido en un hiper-rectángulo de la misma clase a la cual cambia el dato 131
2.3.2. El dato está incluido en un hiper-rectángulo que representa a otra clase distinta 131
2.4. Sub-clasificando muestras 132
2.5. Realizando varios cambios simultáneamente 133
3. Actualizando reglas de clasificación 135
4. Intervención del experto 136
5. Análisis de rendimiento 137
5.1. Costo en hallar el hiper-rectángulo (u hoja) 139
5.2. Re-estructuración del hiper-rectángulo (u hoja) 139
5.3. Conclusiones 140

Capítulo 4. Resultados y Comparaciones 143
1. Ejemplos ficticios en 2D 144
1.1. Configuración de la estrategia 145
1.2. Dos clases separadas 146
1.2.1. Descripción del ejemplo 146
1.2.2. Resultado 146
1.3. Una clase entremedio de otra 147
1.3.1. Descripción del ejemplo 147
1.3.2. Resultado 147
1.4. Una clase envolviendo parcialmente a otras dos 148
1.4.1. Descripción del ejemplo 148
1.4.2. Resultado 149
1.5. Envolturas sucesivas 150
1.5.1. Descripción del ejemplo 150
1.5.2. Resultado 150
1.6. Tres clases con varias zonas de superposición 152
1.6.1. Descripción del ejemplo 152

1.6.2. Resultado 152
1.7. Doble espiral 154
1.7.1. Descripción del ejemplo 154
1.7.2. Resultado 154
1.8. Una clase que encierra a otra 156
1.8.1. Descripción del ejemplo 156
1.8.2. Resultado 156
1.9. Una clase que encierra a otra de manera más ajustada 157
1.9.1. Descripción del ejemplo 157
1.9.2. Resultado 157
1.10. División en diagonal 158
1.10.1. Descripción del ejemplo 158
1.10.2. Resultado 158
1.11. Dos clases compartiendo un sector del espacio 159
1.11.1. Descripción del ejemplo 159
1.11.2. Resultado 160
1.12. Mezcla total de dos clases 161
1.12.1. Descripción del ejemplo 161
1.12.2. Resultado 161
1.13. Resumen 162
2. Bases de datos del repositorio UCI 163
2.1. Bases de datos usadas 165
2.1.1. Ecoli data set 165
2.1.2. Glass data set 165
2.1.3. Haberman's Survival data set 165
2.1.4. Image segmentation data set 165
2.1.5. Ionosphere data set 166
2.1.6. Iris data set 166
2.1.7. Liver disorders data set 166
2.1.8. Pima indians diabetes data set 166
2.1.9. Connectionist bench (Sonar, mines vs. rocks) data set 166
2.1.10. Statlog (Vehicle silhouettes) data set 167
2.1.11. Connectionist bench
(Vowel recognition – Deterding data) data set 167
2.1.12. Wine data set 167
2.1.13. Breast cancer Wisconsin (Original) data set 167
2.1.14. Forest Covertype data set 167
2.2. Resultados 167
3. Comparaciones con otros métodos 169
3.1. C4.5 169
3.2. EHS-CHC 171
3.3. PSO/ACO2 171

3.4. Resultados 172
3.5. Análisis de rendimiento 179
3.5.1. C4.5 179
3.5.2. EHS-CHC 180
3.5.3. PSO/ACO2 181
3.5.4. Resultados 183
4. Minería incremental 184

Capítulo 5. Discusión y trabajo a futuro 187
1. CLUHR 188
1.1. Índices de separabilidad 188
1.2. Supervisión de un experto en el dominio del problema 189
1.3. Adaptabilidad 190
1.4. Comparaciones 191
1.5. Trabajando con valores decrecientes para μ 192
2. Trabajo a futuro 194
2.1. CLUHR mejorado 196
2.1.1. Índices 196
2.1.2. Unión de hiper-rectángulos 197
2.1.3. Simplificación de reglas 198
2.1.4. Operaciones con otros dominios de datos 201
2.1.5. Implementación de una herramienta de supervisión para expertos 201

Bibliografía 202