Inteligencia Artificial
Grado en Ingeniería Informática en Sistemas de Información
Curso 2019/20
HOMEWORK #5: Búsqueda Local

Con la revolución digital, capturar información es fácil y almacenarla es extremadamente barato. Pero ¿para qué se almacenan los datos? Además de la facilidad y conveniencia, se almacenan datos porque se piensa que son un activo valioso en sí mismos. Para los científicos los datos representan observaciones cuidadosamente recogidas de algún fenómeno en estudio; en los negocios, los datos guardan informaciones sobre mercados, competidores y clientes; en procesos industriales recogen valores sobre el cumplimiento de objetivos; etc.

Sin embargo, en general, los datos en bruto raramente son provechosos. Su verdadero valor radica en la posibilidad de extraer información útil para la toma de decisiones o la exploración y comprensión de los fenómenos que dieron lugar a los datos. Una de las principales tareas que se realiza sobre los datos es la de **Selección de Atributos**, que consiste en extraer un subconjunto de atributos o características relevantes de manera que no se pierda capacidad de extraer conocimiento.

La selección de atributos se puede considerar como un problema de búsqueda en un cierto espacio de estados, donde cada estado corresponde con un subconjunto de atributos, y el espacio engloba todos los posibles subconjuntos que se pueden generar. El proceso de selección de atributos puede entenderse como el recorrido de dicho espacio hasta encontrar un estado (combinación de atributos) que optimice alguna función definida sobre un conjunto de atributos. En general, un algoritmo de selección consta de dos componentes básicos: medida de evaluación y método de búsqueda.

Todavía en la actualidad, el evaluar todas las combinaciones de atributos posibles de una base de datos es intratable, incluso aquellas con un número reducido de atributos, debido al coste computacional asociado a la operación. Por ejemplo, en una base de datos con veinte atributos, el número de evaluaciones posibles supera el millón, si se intentara aplicar a una base de datos con miles de atributos podría ser interminable. En este trabajo práctico se debe utilizar como método de búsqueda un algoritmo genético, y como medida de evaluación, CFS, una función que se basa en correlaciones, de manera que considera que los atributos que componen un subconjunto deben estar muy correlacionados con la clase y poco correlacionados entre ellos (Ver Apéndice 1).

Los datos de entrada se muestran en el Apéndice 2.

Implementar las clases java del entorno del problema de selección de atributos tomando como ejemplo el problema de las n-Reinas de la librería AIMA:

1. `SelAttBoard`: representa un subconjunto de atributos, indicando con unos y ceros si cada atributo está o no presente en el subconjunto.
2. `SelAttGenAlgoUtil`: clase de utilidad para el uso de algoritmos genéticos en el entorno de selección de atributos.
3. `SelAttGoalTest`: devolverá siempre falso al no saber de antemano el mejor subconjunto.
4. Crear en el mismo paquete una copia de la clase `GeneticAlgorithm` con el nombre `GeneticAlgorithmSelAtt`, de manera que:
    a. El mejor individuo de la población pase a la siguiente generación.
    b. Usar el cruce uniforme, de manera que se escoge aleatoriamente si cada gen del hijo se toma del primer o del segundo padre.
    c. Incluir también, un operador de aceptación, que después de realizar el cruce y la mutación de los individuos de la población decida si aceptamos los hijos generados. La técnica habitual es la de "Aceptación total", donde todos los hijos generados son aceptados y pasan a formar parte de la nueva población. En este caso se pide implementar la "Aceptación de mejora", donde los hijos pasan a la nueva población si son mejores que el peor de los individuos de la población actual.

5. Incluir en este mismo paquete una clase denominada `SelAttDemo` para probar la implementación anterior. Realizar varias ejecuciones con diferentes valores para los parámetros del algoritmo evolutivo, como la probabilidad de mutación, el número de individuos de la población y el tiempo de ejecución. Observar qué atributos relevantes se obtiene dependiendo de los valores de los parámetros usados y qué valor tiene la medida de evaluación en cada caso.

Apéndice 1: función de evaluación CFS

CFS (Correlation–based Feature Selection) intenta obtener el conjunto de atributos más correlacionado con la clase y con menos correlación entre sí. Se le puede asociar con distintas técnicas de búsqueda. A continuación se puede ver la función de evaluación en rojo, y a la derecha varios ejemplos de su aplicación a distintos subconjuntos de atributos: el resultado de evaluar x_3 es 0,185; el de evaluar el subconjunto $\{x_3,x_1\}$ es 0,220; y con el subconjunto $\{x_3,x_1,x_4\}$ CFS obtiene el valor 0,226. La tabla de correlaciones que aparece abajo recoge la correlación lineal entre los distintos atributos y entre los distintos atributos y la clase.

Correlaciones (CFS)

$$\mathcal{E}(\mathcal{S}) = \frac{k \times \overline{r_{cl}}}{\sqrt{k + k \times (k-1) \times \overline{r_{at}}}}$$

$\mathcal{S}$ → subconjunto de atributos
k → número de atributos
$\overline{r_{cl}}$ → media correlación clase
$\overline{r_{at}}$ → media corr. entre atts.

Correlación	x_1	x_2
x_1	1,000	0,116
x_2	0,116	1,000
x_3	0,022	0,248
x_4	0,007	0,028
Clase	0,130	0,025

	x_3	x_4
x_1	0,022	0,007
x_2	0,248	0,028
x_3	1,000	0,000
x_4	0,000	1,000
Clase	0,185	0,081

Ejemplo de selección con SF+CFS

Subc. $\mathcal{S}$	k	$\overline{r_{cl}}$	$\overline{r_{at}}$	$\mathcal{E}(\mathcal{S})$
$\emptyset$	0	N/D	N/D	0,000
x_1	1	0,130	1,000	$\frac{1\times0,130}{\sqrt{1+1\times(1-1)\times1,000}} = 0,130$
x_2	1	0,025	1,000	$\frac{1\times0,025}{\sqrt{1+1\times(1-1)\times1,000}} = 0,025$
x_3	1	0,185	1,000	$\frac{1\times0,185}{\sqrt{1+1\times(1-1)\times1,000}} = 0,185$
x_4	1	0,081	1,000	$\frac{1\times0,081}{\sqrt{1+1\times(1-1)\times1,000}} = 0,081$
x_3, x_1	2	0,158	0,022	$\frac{2\times0,158}{\sqrt{2+2\times(2-1)\times0,022}} = 0,220$
x_3, x_2	2	0,105	0,258	$\frac{2\times0,105}{\sqrt{2+2\times(2-1)\times0,258}} = 0,133$
x_3, x_4	2	0,133	0,000	$\frac{2\times0,133}{\sqrt{2+2\times(2-1)\times0,000}} = 0,188$
x_3, x_1, x_2	3	0,113	0,132	$\frac{3\times0,113}{\sqrt{3+3\times(3-1)\times0,132}} = 0,175$
x_3, x_1, x_4	3	0,132	0,009	$\frac{3\times0,132}{\sqrt{3+3\times(3-1)\times0,009}} = 0,226$
x_3, x_1, x_4, x_2	4	0,105	0,072	$\frac{4\times0,105}{\sqrt{4+4\times(4-1)\times0,072}} = 0,191$

Apéndice 2: datos de entrada

Se utilizará como entrada el conjunto de datos Breast Cancer, que contiene las características de pacientes para la predicción del cáncer de mama, formado por 286 instancias, 9 atributos y una clase binaria. A continuación, se muestra la tabla de correlaciones donde se indican las correlaciones entre los diferentes atributos, y entre los atributos y la clase:

	A1	A2	A3	A4	A5	A6	A7	A8	A9
A1	1	0.64491	0.65459	0.48636	0.52182	0.58730	0.55843	0.53583	0.35003
A2	0.64491	1	0.90688	0.70558	0.75180	0.68680	0.75572	0.72286	0.45869
A3	0.65459	0.90688	1	0.68308	0.71967	0.70961	0.73595	0.71945	0.43891
A4	0.48636	0.70558	0.68308	1	0.59960	0.66505	0.66672	0.60335	0.41763
A5	0.52182	0.75180	0.71967	0.59960	1	0.58126	0.61610	0.62888	0.47910
A6	0.58730	0.68680	0.70961	0.66505	0.58126	1	0.67590	0.57736	0.33874
A7	0.55843	0.75572	0.73595	0.66672	0.61610	0.67590	1	0.66588	0.34417
A8	0.53583	0.72286	0.71945	0.60335	0.62888	0.57736	0.66588	1	0.42834
A9	0.35003	0.45869	0.43891	0.41763	0.47910	0.33874	0.34417	0.42834	1
Clase	0.71600	0.81790	0.81893	0.69680	0.68278	0.81605	0.75662	0.71224	0.42317