



Universidad Simón Bolívar
Departamento de Computación y T.I.
Inteligencia Artificial II
Septiembre-Diciembre 2009.

TAREA 2 (15 PUNTOS)

Se desea que utilice un algoritmo genético (GA) para predecir si una persona gana más de \$50.000 anualmente. Se le suministrará un conjunto de datos con información socio-económica de más de 48 mil personas que podrá utilizar para entrenar y evaluar su solución.

Su implementación del GA debe partir del sistema GABIL [1]. Una breve descripción de este sistema también la pueden conseguir en el capítulo 9 del Mitchell [3]. Adicionalmente, se espera lo siguiente de su proyecto:

- Deberá codificar y probar 3 versiones de las funciones de selección y recombinación. Una de las funciones de selección deberá ser la de “rueda de ruleta”. En [2] podrá encontrar la descripción de varios operadores.
- Debe incluir las dos extensiones propuestas al algoritmo básico de GABIL: el operador “AddAlternative” y el operador “DropCondition”.

Debe realizar experimentos para conseguir la mejor configuración (combinación de los operadores de selección y recombinación). Una vez determinada la mejor configuración, para la mejor de estas configuraciones realizar variaciones sobre las tasas de mutación y crossover (al menos 3 valores para cada una). Sobre la mejor configuración y parámetros hallados, probar con cada uno de los operadores de extensión (“AddAlternative” y “DropCondition”) por separado y con los dos operadores juntos.

El conjunto de datos está compuesto por 8 atributos categóricos, 6 atributos enteros y la clasificación correcta para cada ejemplo. Los datos y su descripción los puede conseguir en <http://archive.ics.uci.edu/ml/datasets/Adult>

Para simplificar la representación de los atributos enteros, puede dividir el rango de valores posibles en varios sub-segmentos y considerar cada uno como una categoría. Por ejemplo, si el atributo entero X toma valores entre 0 y 5000, puede considerar utilizar las categorías $C_1: 0 \leq X < 2000$, $C_2: 2000 \leq X < 4000$ y $C_3: 4000 \leq X \leq 5000$. Nótese que las categorías no tienen que ser del mismo tamaño.

Pueden implementar su propia librería de algoritmos genéticos o utilizar cualquiera ya existente. Pueden trabajar en el lenguaje de programación de su preferencia.

ENTREGA

La fecha de entrega jueves 19 de noviembre a la 1:30 pm en el laboratorio del Grupo de Inteligencia Artificial. Deberá entregar:

- Impreso:
Un breve informe que contenga:

1. La descripción de la implementación o la librería que utilizó

2. La descripción del AG (los parámetros base) que usaron.
3. El descripción y análisis de los experimentos realizados.
4. En el informe deben dar respuesta a las siguientes preguntas:
 - a) ¿Cuál es la mejor configuración de su algoritmo genético para clasificar los datos estudiados?
 - b) ¿Son útiles los operadores de la extensión?
 - c) ¿Cuál es el mejor conjunto de reglas hallado por el algoritmo genético? Considere el número de ejemplos clasificados correcta e incorrectamente.

Recuerde que por ser los algoritmos genéticos algoritmos estocásticos deben reportar el promedio de varias corridas (al menos 10) para cada configuración.

■ Por correo:

Deberán enviar un correo a julio@cbm.usb.ve con un archivo `tar.gz` que contenga el proyecto y el informe en PDF.

Referencias

- [1] Kenneth A. De Jong, William M. Spears, and Diana F. Gordon. Using genetic algorithms for concept learning. *Mach. Learn.*, 13(2-3):161–188, 1993.
- [2] Melanie Mitchell. Handbook of genetic algorithms (l. d. davis). *Artif. Intell.*, 100(1-2):325–330, 1998.
- [3] Tom M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997.