

Proyecto 2

Fecha de entrega: Martes 12 de Marzo

Se requiere que, usando un algoritmo genético (GA), se realice la clasificación un conjunto de datos correspondiente a la aprobación de créditos de una entidad financiera.

El problema es predecir en los ejemplos dados si a un individuo se le otorga un crédito o no.

La descripción de los datos, así como los conjuntos ejemplos (que deben separar en los conjuntos de entrenamiento y prueba) se encuentran disponibles en:

<http://archive.ics.uci.edu/ml/datasets/Credit+Approval> . Este conjunto de datos está compuesto por 690 ejemplos con 6 atributos categóricos, 3 atributos binarios, y 6 continuos (enteros o reales) y la clasificación correcta para cada ejemplo.

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 [2]. Pueden usar cualquier librería de GAs en el lenguaje de su preferencia.

Se espera lo siguiente de su proyecto:

- Deberán codificarse y probarse 2 versiones de la función de selección de padres y 2 versiones de la función de selección sobrevivientes. Una de las funciones de selección deberá ser la de “rueda de ruleta”.
- Agregar las dos extensiones propuestas al algoritmo básico de GABIL: el operador “AddAlternative” y el operador “Drop Condition”.
- Incorporar a su función de fitness un mecanismo para controlar el tamaño de los clasificadores.

Deben realizar experimentos para conseguir la mejor configuración (combinación de los operadores de selección de padres y sobrevivientes). 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 “Drop Condition”) por separado y con los dos operadores juntos. Probar variaciones sobre la penalización al tamaño de los clasificadores y comparar los resultados obtenidos al incluir o no dicha penalización.

Nota: Para simplificar la representación de los atributos continuos , pueden 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: $C1 : 0 \leq X \leq 2000$, $C2 : 2000 \leq X \leq 4000$ y $C3 : 4000 \leq X \leq 5000$. Nótese que las categorías no tienen que ser del mismo tamaño.

Entrega: La fecha de entrega Martes 12 de Marzo (1:30 pm). Deberán entregar:

- Impreso:
 - Un breve informe que contenga:
 1. La descripción de la implementación (o uso de librería).
 2. La descripción del AG (los parámetros base) que usaron.
 3. La descripción y el 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? Considerando el número de ejemplos clasificados correcta e incorrectamente.
- d) Describa la función de fitness utilizada. Es útil incluir en la función de fitness un factor de penalización a clasificadores muy grandes?

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

- En el aula virtual:

Deberán subir un archivo tar.gz que contenga el código del proyecto, el informe en PDF, e instrucciones de ejecución del programa (README.txt).

Referencias

- [1] Kenneth A. De Jong and William M. Spears and Diana F. Gordon. *Using Genetic Algorithms for Concept Learning*, Machine Learning, 13, pp. 161-188, 1993, <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.48.6461>. consultado el 22/02/2013.
- [2] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.
- [3] Melanie Mitchell, *Handbook of genetic algorithms*. Artif. Intell., 100(1-2): 325–330, 1998.