PRÁCTICA 2. ÁRBOL DE DECISIÓN (ID3)

INGENIERÍA DEL CONOCIMIENTO

Mª Victoria Barylak Alcaraz

Detalles de implementación.

Para la realización de la segunda práctica de la asignatura se ha implementado el algoritmo *ID3* usando como lenguaje de programación *Java* y, como entorno de programación, *Eclipse Neon*.

El algoritmo *ID3* se ha implementado como una función recursiva que recibe como parámetros un nodo del árbol de decisión que se va a construir, así como la lista de atributos y ejemplos en forma de tabla.

En cada iteración, el algoritmo comprueba si todos los ejemplos recibidos por parámetro son positivos, en cuyo caso devolverá un nodo positivo, si todos los ejemplos son negativos se devolverá un nodo negativo y, en caso de no cumplirse ninguna de las dos premisas anteriores, se procede al cálculo de los méritos de los atributos.

Para esto se calcula, para cada atributo y su respectivo conjunto de valores, el número de ejemplos totales en el que aparece cada valor, el número de ejemplos negativos, y el número de ejemplos positivos. Una vez calculado todo eso se utiliza la siguiente fórmula para el cálculo del mérito de un atributo:

$$mérito = \frac{r_i}{N} * infor(n_i, p_i)$$

Donde

$$infor(p, n) = -p * \log_2 p - n * \log_2 n$$

Y $n = \frac{nº\ ejemplos\ negativos}{nº\ ejemplos\ totales}$, $p = \frac{nª\ ejemplos\ positivos}{nª\ ejemplos\ totales}$, y

$\frac{r_i}{N} = nº\ de\ ejemplos\ con\ valor\ i\ entre\ el\ nº\ total\ de\ ejemplos.$

En el cálculo de la entropía ($infor$) se comprueba que ni n ni p sean iguales a 0, ya que eso daría lugar a un error.

Una vez calculados los méritos de todos los atributos se escoge el atributo con menor mérito como el nodo raíz y se sigue con la siguiente llamada recursiva, esta vez con el hijo de la raíz como nodo, y con la tabla debidamente reestructurada (eliminando la columna del atributo elegido, y seleccionando sólo los ejemplos donde aparezca el valor que se esté examinando en ese momento).

La primera ampliación que se ha realizado es la de implementar todos los niveles de recursión de modo que, dado un archivo con la lista de atributos y otro con la lista de ejemplos, el programa es capaz de mostrar por pantalla las reglas deducidas del árbol de decisión.

Otra ampliación que se ha llevado a cabo es la opción de introducir cualquier lista de atributos y ejemplos para la deducción de reglas, siempre y cuando estén en el formato adecuado (el cuál se explica en el manual de usuario).

El proyecto no cuenta con interfaz gráfica, tanto la interacción con el usuario como la muestra de las reglas deducidas se lleva a cabo por consola.