MINERIA DE DATOS UTILIZANDO SISTEMAS INTELIGENTES

PRACTICA 2 – ARBOLES

Material de Lectura:

Capítulo 11 del Libro Introducción a la Minería de Datos de Hernández Orallo

Ejercicio 1

a) Construya manualmente, a partir de los datos la hoja **Train** del archivo **preingreso.xls** y de la medida de Desorden Promedio vista en clase, el árbol de clasificación capaz de predecir si un alumno aprobará o no el examen. Indique en cada paso los valores de desorden obtenidos y las selecciones realizadas. Dibuje y explique el árbol obtenido. Sugerencia: se puede utilizar el operador *"Weight by Information Gain"* de Rapidminer para facilitar los cálculos.

   ¿Podría recomendar, en base al árbol obtenido, alguna medida a tomar para mejorar la tasa de aprobados?

b) Utilice el operador J48 de Rapid Miner para obtener el árbol de clasificación solicitado en a)

   Analice la precisión del árbol utilizando los operadores *"ApplyModel"* y *"Performance"* sobre los datos de la hoja **Train**. Incluya en su informe la matriz de confusión correspondiente y explique qué significan los valores que aparecen en ella.

c) Utilizando el árbol obtenido en b), indique los valores correspondientes a la columna *"Examen_Final"* de la hoja **Test** del archivo **preingreso.xls**. Realice este proceso en forma manual.

d) Indique si es posible utilizar Rapid Miner para clasificar los ejemplos de la hoja **Test**. Si su respuesta es afirmativa, detalle el proceso realizado. Note que el resultado de esta operación es equivalente a lo solicitado en c).

Ejercicio 2

Aplique el operador ID3 de RapidMiner a los datos de entrenamiento del archivo **preingreso.xls** (hoja **Train**) realizando dos tipos distintos de discretización de los atributos numéricos:

a) "Discretizebybins" utilizando tres intervalos
b) "Discretizebyuserspecification" utilizando los siguientes intervalos:
   a. "Asistencia" (0 a 3→ Baja, 4 a 6→ Media y 7 a 10 →Alta)
   b. "Entregas_Completas" (0 a 4 → Pocas, 5 a 7→ Suficientes y 8 a 10 → Muchas)

Clasifique utilizando Rapid Miner los ejemplos de la hoja Test para cada uno de los casos (a y b). ¿Se obtuvo elmismo resultado? ¿Por qué razón?

Ejercicio 3

a) Utilice los datos del archivo **DrogasNominalTrain.xlsx** para construir un árbol de decisión ID3 utilizando la medida de desorden "information gain", con "minimal size for split" igual a 4 y "minimal leaf size" igual a 2. Observe la performance obtenida sobre el conjunto de entrenamiento. Observe el tamaño y la complejidad del árbol resultante.

b) Aplique el árbol obtenido en a) sobre los datos del archivo **DrogasNominalTest.xlsx**. Observe la performance obtenida.

c) Repita los pasos a) y b) utilizando "minimal size for split" igual a 15 y "minimal leaf size" igual a 10. Observe nuevamente la complejidad del árbol resultante y la performance obtenida en los conjuntos de entrenamiento y test.

d) Repita los pasos a) y b) utilizando "minimal size for split" igual a 30 y "minimal leaf size" igual a 20. Observe nuevamente la complejidad del árbol resultante y la performance obtenida en los conjuntos de entrenamiento y test.

e) ¿Qué ocurre con el tamaño del árbol en cada caso? ¿Se observan diferencias en la performance del árbol en ambos conjuntos de datos? ¿A qué se deben estas diferencias?