OCR-Weka: Árboles de clasificación (DT) - J48

NOMBRE: Frédéric Roux **Fecha: 13/12/2018**

En la carpeta *classifiers/trees* de la pestaña *Classify* de Weka puedes encontrar el clasificador *J48*. Se trata de la implementación Weka del conocido clasificador *C4.5*. El algoritmo *J48* se ajusta al algoritmo *C4.5* al que se le amplían funcionalidades tales como permitir la realización del proceso de podado mediante *reducedErrorPruning* (el conjunto de ejemplos se divide en un subconjunto de entrenamiento y otro de test que se utiliza para estimar el error de la poda) o que las divisiones sean siempre binarias (*BinarySplits*).

El parámetro más importante es el que tiene que ver con la complejidad del clasificador construido, y que en el caso de J48, es el parámetro *confidence factor*, que controla el tamaño del árbol de decisión construido. No se puede controlar directamente el número de nodos, pero cuanto más pequeño es este parámetro, mas simple tiende a ser el clasificador (menos nodos), y viceversa. Este parámetro varía entre 0 y 1, siendo el valor por omisión de 0.25. La ventaja de un modelo más simple, siempre que sea lo suficientemente complejo para que el aprendizaje sea correcto, es que es más sencillo de entender. Este parámetro tiene efecto siempre que se mantenga la opción de poda activada (*unpruned=False*).

Otro parámetro que interviene en el desarrollo del árbol es el número de instancias mínimo por nodo hoja (*nimNumObj*). Este número indica el número mínimo de ejemplos que debe haber por cada uno de los dos nodos hoja que resultarían de la división de un nodo, con lo que no se considerarían divisiones que no cumplieran ese tamaño mínimo.

Un árbol se puede mostrar en Weka de dos formas: (a) formato gráfico, para lo que hay que seleccionar la opción *Visualize Tree* pinchando con el botón derecho sobre la ventana *Result-list* (se puede ajustar el árbol gráfico al tamaño de la pantalla de visualización con la opción *Fit to screen*); (b) formato ASCII, en la parte derecha del interfaz de Weka (*Classifier output*).

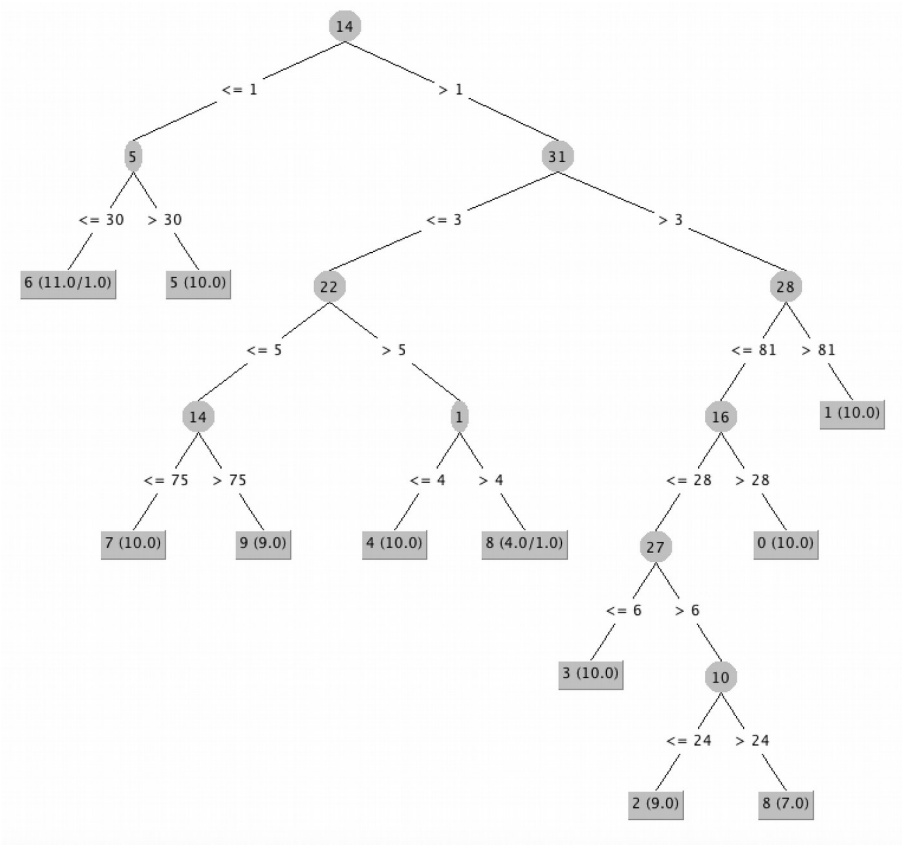
Responde las siguientes preguntas:

(a) ¿Cuál es la tasa de acierto obtenida para el ejemplo de OCR (muestra de test) con el árbol podado? Incluye en este documento el árbol gráfico obtenido. Recuerda entrenar el clasificador con la muestra de entrenamiento.

La tasa de acierto obtenida para el ejemplo OCR es de 55%.

```
14 <= 1
|   5 <= 30: 6 (11.0/1.0)
|   5 > 30: 5 (10.0)
14 > 1
|   31 <= 3
|   |   22 <= 5
|   |   |   14 <= 75: 7 (10.0)
|   |   |   14 > 75: 9 (9.0)
|   |   22 > 5
|   |   |   1 <= 4: 4 (10.0)
```

```
|   |   |   1 > 4: 8 (4.0/1.0)
|   31 > 3
|   |   28 <= 81
|   |   |   16 <= 28
|   |   |   |   27 <= 6: 3 (10.0)
|   |   |   |   27 > 6
|   |   |   |   |   10 <= 24: 2 (9.0)
|   |   |   |   |   10 > 24: 8 (7.0)
|   |   |   16 > 28: 0 (10.0)
|   |   28 > 81: 1 (10.0)
```



(b) ¿Qué indican los nodos y los arcos? ¿Qué crees que pueden ser los números que aparecen en las hojas? Para pensar sobre ello, ten en cuenta cuántos casos de cada dígito hay en el conjunto de entrenamiento. A la vista de esos resultados, ¿cuál será la tasa de acierto para el conjunto de training?

Los nodos : las variables independientes (o predictoras)

Los arcos : los arcos indican el criterio de división mediante el cual se justifica
la división de la sub-población de un nodo del árbol en una serie de estratos que formarán los nodos hijos respecto a la variable dependiente o clase.

Números en las hojas : El primer numero indica el valor predecido de la clase 0-9. El primer valor entre parentesis indica el nombre de casos que tienen un valor predecido de 0-9, y el secundo valor indica el nombre de casos que están mal clasificados. En el conjunto de entrenamiento hay n=10 casos de cada dígito y hay solo 2 casos mal clasificados, por lo tanto la tasa de acierto para el conjunto de training seria 100-2 = 98%.

(c) Un árbol de clasificación es un paradigma que explica los casos clasificados, ¿qué tiene que ocurrir para que un dígito sea clasificado como '8'?