UNIVERSIDAD ICESI
Analítica de Datos

Taller de Clasificación: árbol de decisión humano en WEKA

Objetivo

El propósito de este taller es entender la idea de base del particionamiento automático que realizan los algoritmos de árboles de decisión, al hacerlo de manera manual.

Herramienta de clasificación

Vamos a utilizar una herramienta particular de WEKA que no está en la versión básica de Weka, por lo que hay que instalarla a manera de plugin o paquete. Los firewalls de la universidad impiden utilizar la aplicación de instalación de paquetes de manera directa, por lo que hay que instalarlo de una manera menos inmediata. Siga los siguientes pasos:

1. Descargar https://sourceforge.net/projects/weka/files/weka-packages/userClassifier1.0.3.zip/download, el archivo del clasificador queda en el directorio de descargas (i.e. C:\Users\EDC1147_01\Downloads)
2. Abra Weka. En el menú inicial, escoja "Tools", luego "Package Manager". La aplicación de paquetes se abre.
3. En la parte superior derecha oprima el botón "File/URL". Esto abre un diálogo.
4. Oprima el botón "Browse", escoja el tipo de archivo “Package archive file”, navegue hasta el directorio de descargas y escoja el archivo **userClassifer1.0.3.**zip
5. Reinicie Weka. El clasificador lo encontrará en la pestaña "Classifier", en el repertorio "tree"

Indicaciones:

En este taller se va a trabajar con un conjunto de datos de imágenes que viene incluido en la versión estándar de Weka conjuntamente con otros datasets. Originalmente fue creado por el equipo “Vision Group” de la Universidad de Massachusetts. La descripción se encuentra aquí: https://archive.ics.uci.edu/ml/datasets/Image+Segmentation. Contiene información relacionada con características de regiones de pixeles de las imágenes (cada instancia describe una región de 3x3 pixeles) y el último atributo especifica la clase de objeto o elemento al que corresponde la región de la imagen: brickface, sky, foliage, cement, window, path, grass.

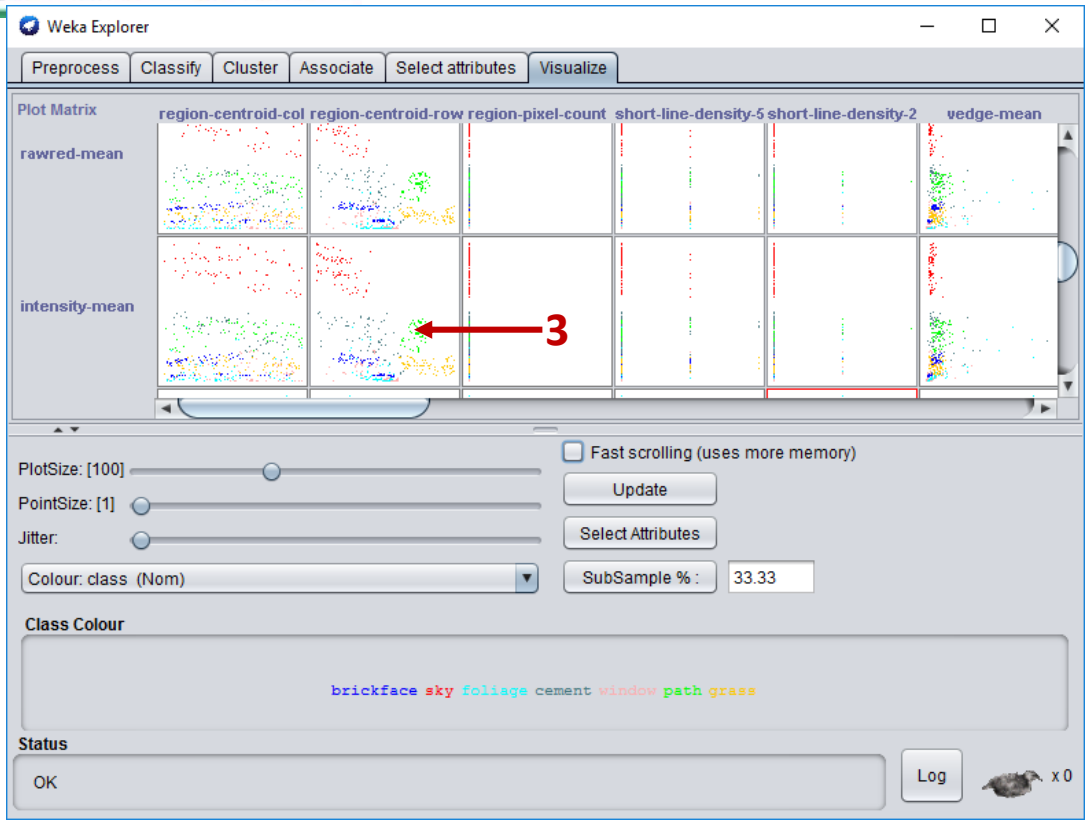
Parte 1: Exploración de datos

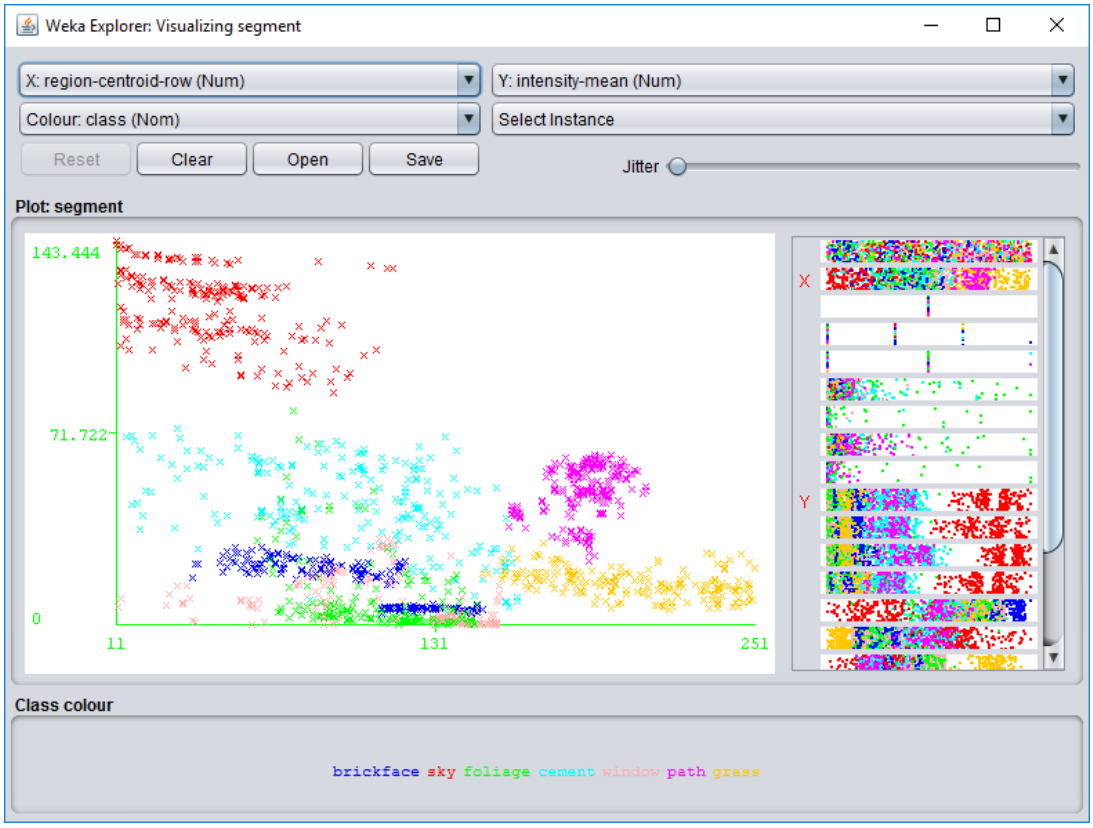
1. Cargue el dataset: En la pestaña “Preprocess”, oprima el botón “Open file…”, navegue hasta el directorio “C:\Program Files\Weka-3-8\data” y escoja el archivo “**segment-challenge.arff**”.

2. Realice un análisis exploratorio rápido de los datos: ¿Cuántos atributos hay? ¿Cuántas instancias? ¿Cuántas clases? ¿Cuántas instancias por clase?

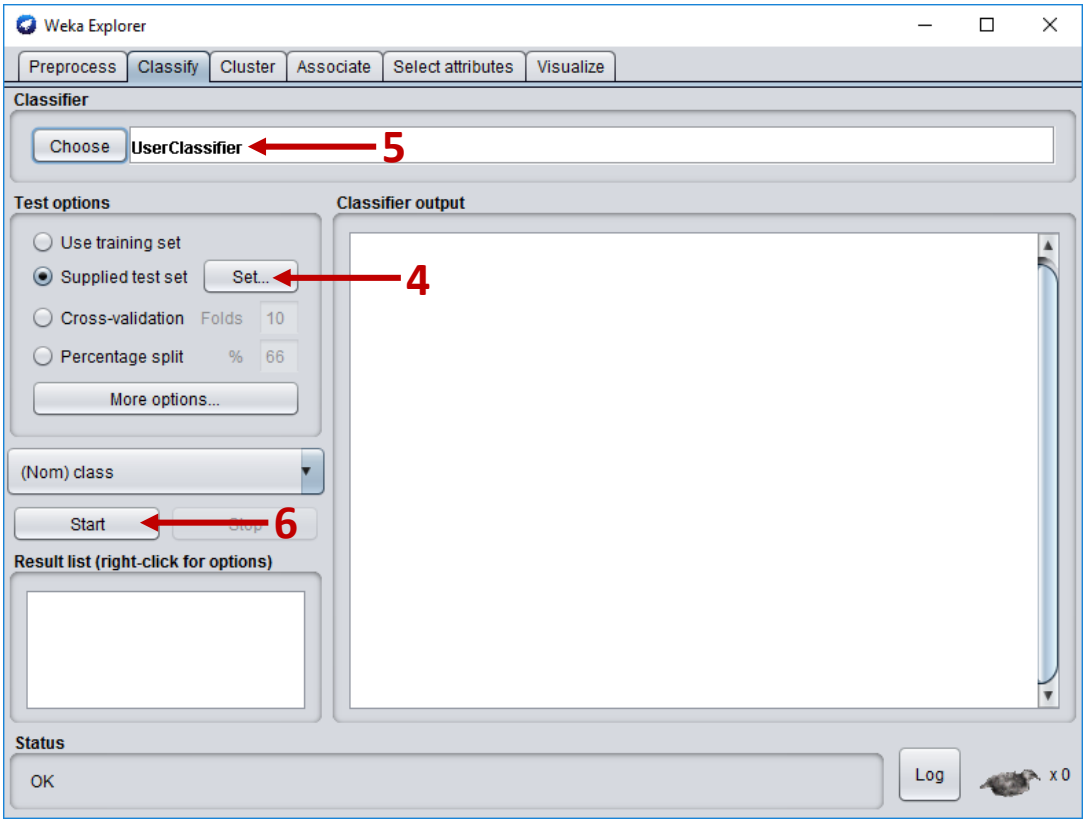
3. En la pestaña “Visualize”, seleccione el plot que confronta el 2º atributo “región-centroid-row” con el 10º atributo “intensity-mean”.

   Esos 2 ejes permiten diferenciar entre las diferentes instancias de las regiones de diferente clase (presentadas en diferentes colores).


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
region-centroid-col
region-centroid-row
region-pixel-count
short-line-density-5
short-line-density-2
vedge-mean
rawred-mean
intensity-mean
3
PlotSize: [100]
PointSize: [1]
Jitter:
Colour: class (Nom)
Fast scrolling (uses more memory)
Update
Select Attributes
SubSample % :
33.33
Class Colour
brickface sky foliage cement window path grass
Status
OK
Log
x 0

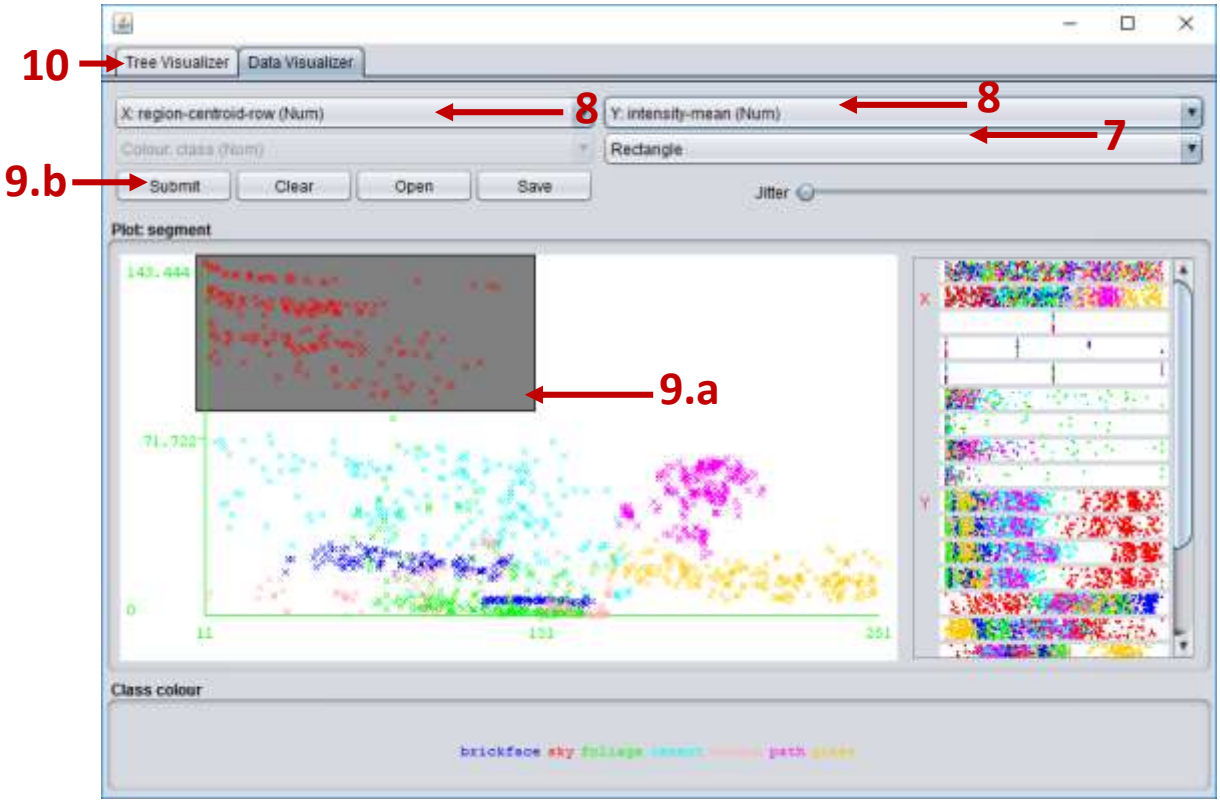

Weka Explorer: Visualizing segment
X: region-centroid-row (Num)
Y: intensity-mean (Num)
Colour: class (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: segment
143.444
71.722
0
11
131
251
X
Y
Class colour
brickface sky foliage cement window path grass

4. Vamos a definir visualmente las regiones “cuadradas” más homogéneas posibles que nos permiten separar la mayoría de las instancias de una misma clase. Vamos a utilizar un archivo de test predefinido, diferente al de entrenamiento. En la pestaña “Classify” En el panel “Test Options”, escoja la opción “Supplied test set”, y oprima el botón “Set”. En el cuadro de diálogo, oprima el botón “Open file…” y escoja el archivo “**segment-test.arff**”, que se encuentra en el mismo repertorio que el anterior dataset de entrenamiento ya cargado.



El atributo correspondiente a la clase es el último, lo que concuerda con los parámetros por defecto de Weka, por lo que no hay que hacer cambios en este respecto.

5. En el panel “Classifer”, oprima el botón “Choose”, y en el repertorio “trees” escoja “UserClassifier”.
6. Oprima el botón “Start”. Esto abrirá una ventana nueva con dos pestañas: “Tree visualizer” y “DataVisualizer”
7. En la pestaña Data Visualizer, definir el tipo de instancia a seleccionar como “Rectangle”



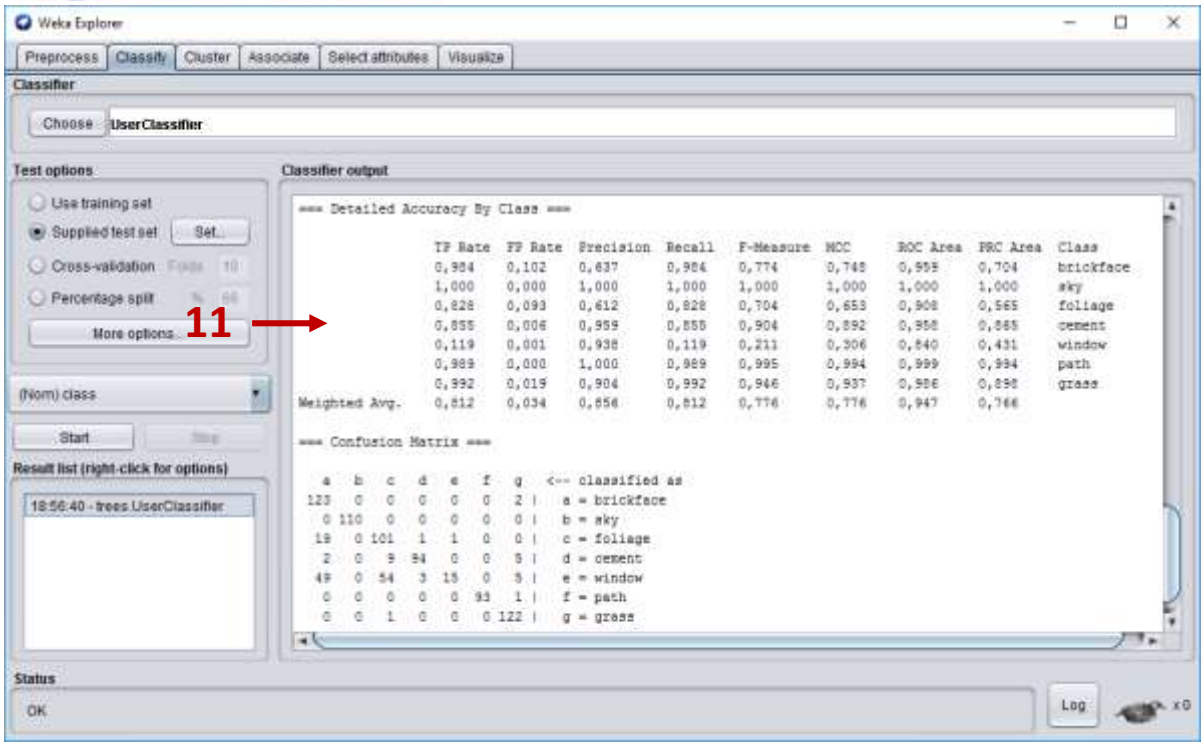
8. Escoja los atributos X y Y que ya visualizamos previamente. Estos presentan una buena separabilidad de las clases del dataset.

9. Vamos a seleccionar una a una varias regiones que a nuestro juicio representen secciones rectangulares que incluyen instancias de una clase determinada con una gran mayoría:

    - Defina la región con el mouse de arriba a la izquierda abajo a la derecha
    - Oprima el botón “Submit”. Esa región será “clasificada” según la clase mayoritaria.
    - Repita hasta que todos los puntos estén clasificados

10. En la pestaña “Tree Visualizer” encontrará el resultado. Analícelo. Presione botón derecho en el canvas con el árbol y escoja la opción “Accept the tree”.

11. El árbol creado va a ser aplicado al dataset de test, y los resultados se muestran en el canvas de la pestaña “Classify del “Explorer” de Weka.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
UserClassifier
Test options
Use training set
Supplied test set
Set...
Cross-validation
Percentage split
More options...
11
(Nom) class
Start
Result list (right-click for options)
18:56:40 - trees.UserClassifier
Classifier output
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0,984 0,102 0,837 0,984 0,774 0,748 0,959 0,704 brickface
1,000 0,000 1,000 1,000 1,000 1,000 1,000 1,000 sky
0,828 0,093 0,612 0,828 0,704 0,653 0,908 0,565 foliage
0,855 0,006 0,959 0,855 0,904 0,892 0,958 0,865 cement
0,119 0,001 0,938 0,119 0,211 0,306 0,840 0,431 window
0,989 0,000 1,000 0,989 0,995 0,994 0,999 0,994 path
0,992 0,019 0,904 0,992 0,946 0,937 0,986 0,898 grass
Weighted Avg. 0,812 0,034 0,856 0,812 0,776 0,776 0,947 0,766
=== Confusion Matrix ===
a b c d e f g <-- classified as
123 0 0 0 0 0 2 | a = brickface
0 110 0 0 0 0 0 | b = sky
19 0 101 1 1 0 0 | c = foliage
2 0 9 94 0 0 5 | d = cement
49 0 54 3 15 0 5 | e = window
0 0 0 0 0 93 1 | f = path
0 0 1 0 0 0 122 | g = grass
Status
OK
Log
x 0