C4.5

Adolfo Marín Arriaga Juan Carlos López López Luis Rodrigo Rojo Morales

13 de noviembre de 2016

- Evaluación de patrones

  El atributo que asegura que una persona esté en la clase >50k es el capital gain, siempre y cuando sea mayor a 6849. En el caso en que sea menor o igual a esa cantidad se toman en consideración otros atributos, principalmente el estado civil, pues si una persona tiene como estado civil Divorced, Married-spouse-absent, Separated o Widowed caeran en la clase <=50k. En el caso en que tenga estado civil Married-AF-spouse, se depende únicamente del atributo educación para saber en qué clase cae la persona, y analizando el árbol generado, las personas que tienen educación Bachelors, Some-college o Assoc-voc caeran en la clase $> 50k$, y los restantes tipos de educación caeran en <=50k. En el caso en que la persona tenga estado civil Married-civ-spouse, se tienen que verificar diversos atributos para al final obtener la predicción de su clase, esto si su educación es distinta a 9th, pues en este caso caera en la clase <=50k.

- Clasificadores para la comparación

  Para la comparación, los árboles C4.5, es decir, los clasificadores que se usaron son los del archivo AlgoritmoC45.java que es nuestra implementación y el de la biblioteca J48 en Weka, pues de igual forma se planeaba usar alguna biblioteca de R pero la única que encontramos es C50, y ésta solo contiene una mejora de C4.5 (C5.0).

- Primer clasificador

  Usando el de AlgoritmoC45.java para clasificar los elementos que vienen en el archivo AdultPrepoc.csv nos da el siguiente resultado tentativo:

```
age
Proporción Ganancia: 3.4877233013944786E-4

workclass
Proporción Ganancia: 0.0025162619881970616

education
Proporción Ganancia: 0.004128244800339935

marital_status
Proporción Ganancia: 0.02891013158892776

occupation
Proporción Ganancia: 0.003015635008779692

relationship
Proporción Ganancia: 0.021771322487953413

race
Proporción Ganancia: 0.00363937039432976

sex
Proporción Ganancia: 0.028516020587302855

capital_gain
Proporción Ganancia: 0.0021416150612594055

capital_loss
Proporción Ganancia: 0.0018248556212221002

hours_per_week
Proporción Ganancia: 3.1835305591222614E-4

native_country
Proporción Ganancia: 3.7934229343783167E-4
```

- Segundo clasificador

  Usando la biblioteca J48 de Weka para clasificar los elementos que vienen en el archivo AdultDataSetTest.csv nos da el resultado:

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       13914               85.4616 %
Incorrectly Classified Instances      2367               14.5384 %
Kappa statistic                          0.5578
Mean absolute error                      0.2048
Root mean squared error                  0.3272
Relative absolute error                 56.7614 %
Root relative squared error             77.0239 %
Total Number of Instances            16281

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area
                 0.945    0.436    0.875      0.945    0.908      0.568    0.868
                 0.564    0.055    0.759      0.564    0.647      0.568    0.868
Weighted Avg.    0.855    0.346    0.848      0.855    0.847      0.568    0.868
```

- Comparación
  El primer clasificador aún está en la fase de pruebas.
  El segundo clasificador clasificó bien a 13914 personas y mal a 2367 personas (14.04 %).
  El primer clasificador acertó en el ? de los casos.
  El segundo clasificador acertó en el 85.46 % de los casos.