Raúl García Furió

Antonio Grimaldo

Ana María González M

## Using Data Mining Tools WEKA

1. Explain the advantages and disadvantages of writing a program on your own vs using a pre-created suite such as WEKA.

The advantages of writing our own program is that we are able to learn  how the algorithm works. But it consumes time and effort to be developed and tested.

Using WEKA or any other similar tools can be very helpful because they are more complex, they have a much bigger process to be developed. And it can be very helpful to reduce the use of code in your projects. WEKA is also well documented that for us, the developers, it's like a drop of water in the desert.

2. Explain what criteria you followed to choose the data sets for your tree and the WEKA tests.

The data must contain nodes that contain different attributes, and it also has to contain a final “attribute” which value will be based in the other's attributes.
Each attribute must have finite set of states that it can take.

Based in this we selected a films data set that looks like this:

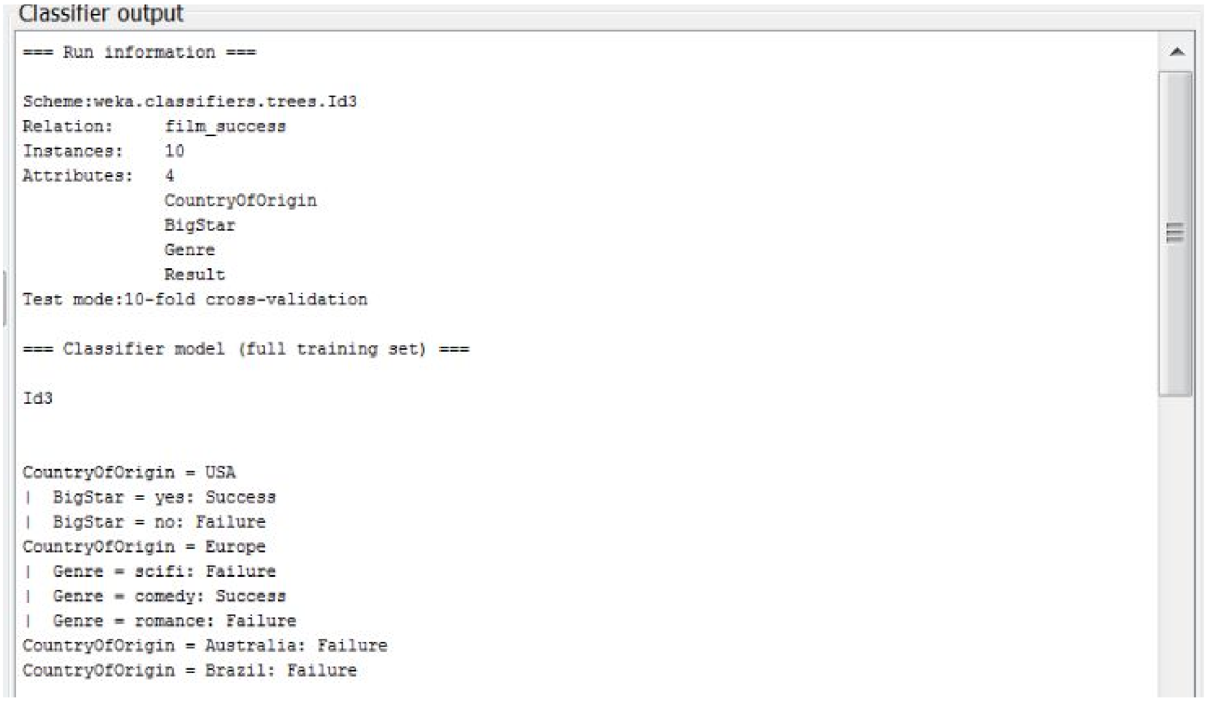
```
@RELATION film_success

@ATTRIBUTE CountryOfOrigin   {USA, Europe, Australia, Brazil}
@ATTRIBUTE BigStar                    {yes, no}
@ATTRIBUTE Genre                      {scifi, comedy, romance}
@ATTRIBUTE Result                     {Success, Failure}

@DATA
USA,yes,scifi,Success
USA,no,comedy,Failure
USA,yes,comedy,Success
Europe,no,comedy,Success
Europe,yes,scifi,Failure
Europe,yes,romance,Failure
Australia,yes,comedy,Failure
Brazil,no,scifi,Failure
Europe,yes,comedy,Success
USA,yes,comedy,Success
```

This set give us a good example because the attribute Genre and Country has more than 2 values, and the rest  has exactly two values

3. Include the graphics of the trees or part of the trees you generated in WEKA and your own program. Are they different, and if so, why?



```
=== Run information ===

Scheme:weka.classifiers.trees.Id3
Relation:     film_success
Instances:    10
Attributes:   4
              CountryOfOrigin
              BigStar
              Genre
              Result
Test mode:10-fold cross-validation

=== Classifier model (full training set) ===

Id3


CountryOfOrigin = USA
|  BigStar = yes: Success
|  BigStar = no: Failure
CountryOfOrigin = Europe
|  Genre = scifi: Failure
|  Genre = comedy: Success
|  Genre = romance: Failure
CountryOfOrigin = Australia: Failure
CountryOfOrigin = Brazil: Failure
```

The format is a little bit different but it makes the same algorithm. It displays first the more complex nodes. In this example if the country is Brazil or Australia is automatically a Failure (sounds like the hollywood people make this data set) and it leave this nodes for the last.

4. Based in what you have learned so far where would you use decision trees?

In a compiler/interpreter, a XML parser for example uses a decision tree.

Many games, like the chess, also uses decision trees. Autocomplete algorithms also use a decision tree