Intelligent systems - Laboratory 5
Bayes Network
02 - 11 - 2018
Fernando Miguel Arriaga Alcántara A01270913
Juan Pablo Ruiz Orantes A01700860

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

First, one great advantage of writing a program by our own is that we can freely control all the hidden parameters, adjusting them depending on the application. In one case we increased the value of entropy needed just to get a proper visualization of the resultant tree. The information gain comparison can also be changed. Another advantage is that in some cases the displayed output can be harder to visualize (i.e. WEKA) when there are too much attributes, and in our script the output is displayed in a list-like information, so it is easier to read.

One advantage from pre-created tools is that they have a very robust parser, avoiding problems in the input like having an extra espace separation or even handle missing values represented by '?' which our implementation can't handle.

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

In general, we wanted the dataset to have no missing values, usually portrayed by '?'. Our program just ignores instances and it does not add up information in those cases. Other thing we looked for in data sets was for them to have as much variety of information as possible, we mean as much combinations as possible of the different values of the attributes. Sometimes, our algorithm when determining information gain would not find an instance of a value in the sub set, therefore, the probability of it appearing would be zero. This means trouble when calculating entropy as it requires the log function of the probability.

Also, we wanted data sets with a small number of values for the attributes in order to have a more readable tree in weka as it tends to expand several different values. As stated before, this is one advantage of our program when a big tree is outputted.

Lastly, we tried to find datasets with interesting information to try and get some interesting conclusions out of our tree.

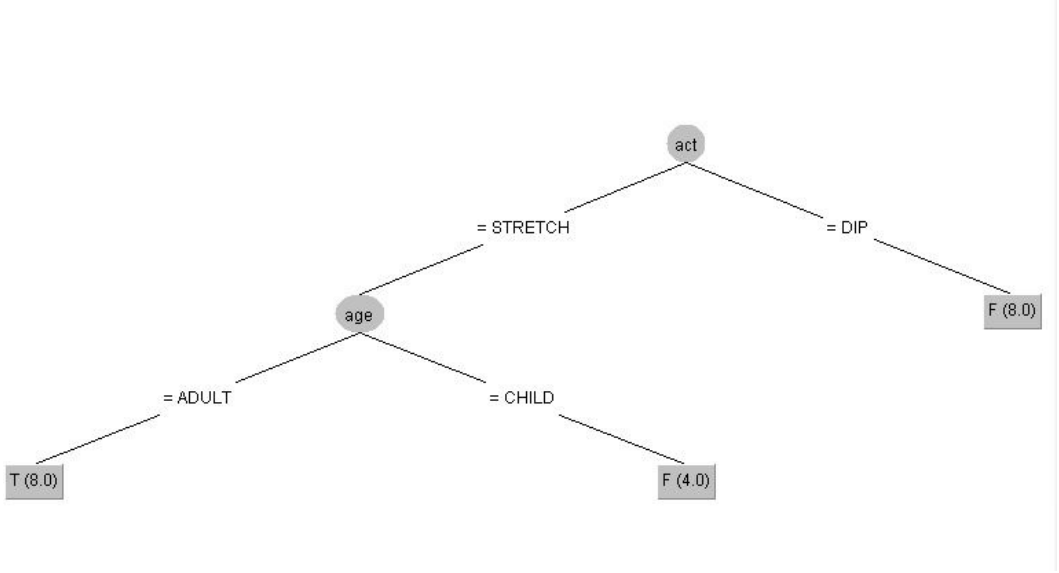
3.- Include the graphics of the trees or part of the trees you generated in WEKA and your own program. Are they different, and if so, why?

The following tree was generated with a dataset from an experiment where participants were shown a set of photographs and asked to predict if the person in the photo would inflate the balloon. The purpose was for them to find the relation between the attributes. We thought this would be interesting to test because our program tries to do just that, tho find the relation between the attributes and the outcome:

The attributes of the data set were:

@relation BreastCancer
@attribute Color {YELLOW, PURPLE}
@attribute size {SMALL, LARGE}
@attribute act {STRETCH, DIP}

@attribute age {ADULT, CHILD}
@attribute inflated {T, F}
The resulting weka tree:



Balloon test weka tree

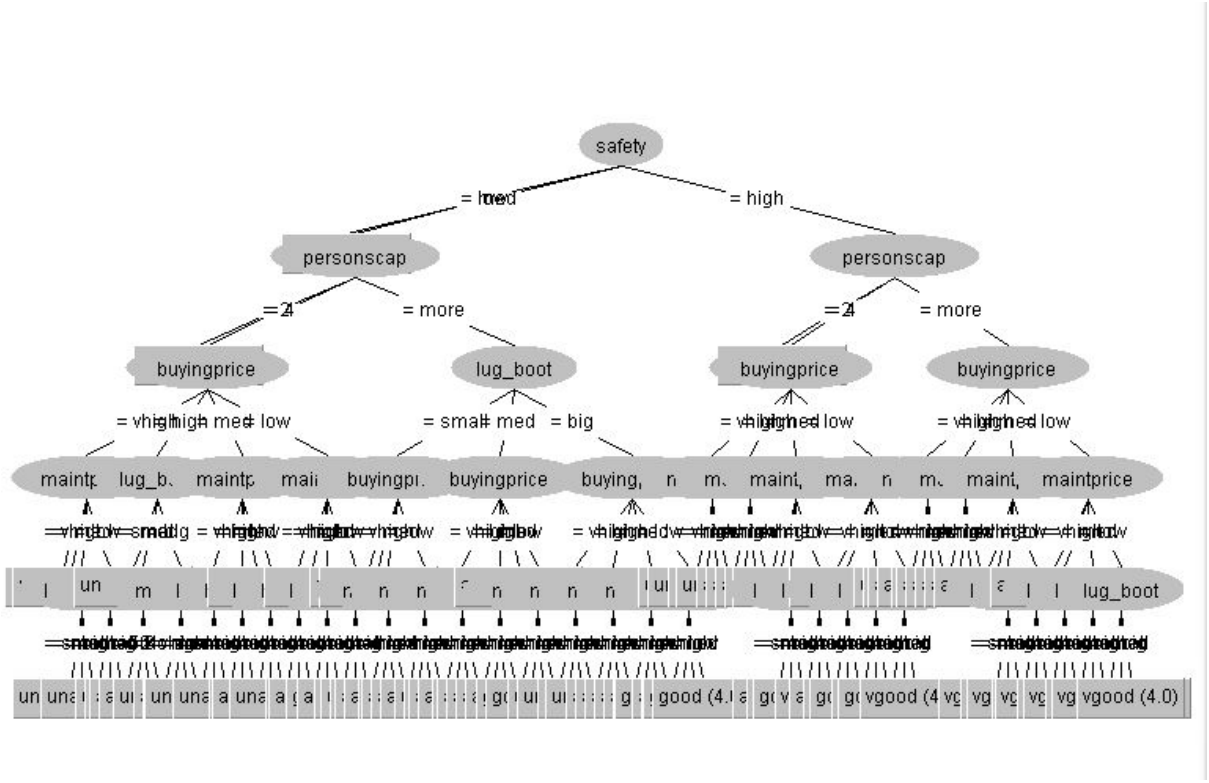
The tree generated by our program:

```
@attribute act    {STRETCH, DIP}
@attribute age    {ADULT, CHILD}
@attribute inflated   {T, F}

@data
YELLOW,SMALL,STRETCH,ADULT,T
YELLOW,SMALL,STRETCH,ADULT,T
YELLOW,SMALL,STRETCH,CHILD,F
YELLOW,SMALL,DIP,ADULT,F
YELLOW,SMALL,DIP,CHILD,F
YELLOW,LARGE,STRETCH,ADULT,T
YELLOW,LARGE,STRETCH,ADULT,T
YELLOW,LARGE,STRETCH,CHILD,F
YELLOW,LARGE,DIP,ADULT,F
YELLOW,LARGE,DIP,CHILD,F
PURPLE,SMALL,STRETCH,ADULT,T
PURPLE,SMALL,STRETCH,ADULT,T
PURPLE,SMALL,STRETCH,CHILD,F
PURPLE,SMALL,DIP,ADULT,F
PURPLE,SMALL,DIP,CHILD,F

act: STRETCH
  age: ADULT
    ANSWER: T
  age: CHILD
    ANSWER: F
act: DIP
  ANSWER: F

Process returned 0 (0x0)   execution time : 12.043 s
```

Balloon test D3

The next data set we tried was used to determine if a car was in an acceptable condition based on several factors.
The attributes of the data set were:
@relation Car condition
@attribute buyingprice {vhigh, high, med, low}

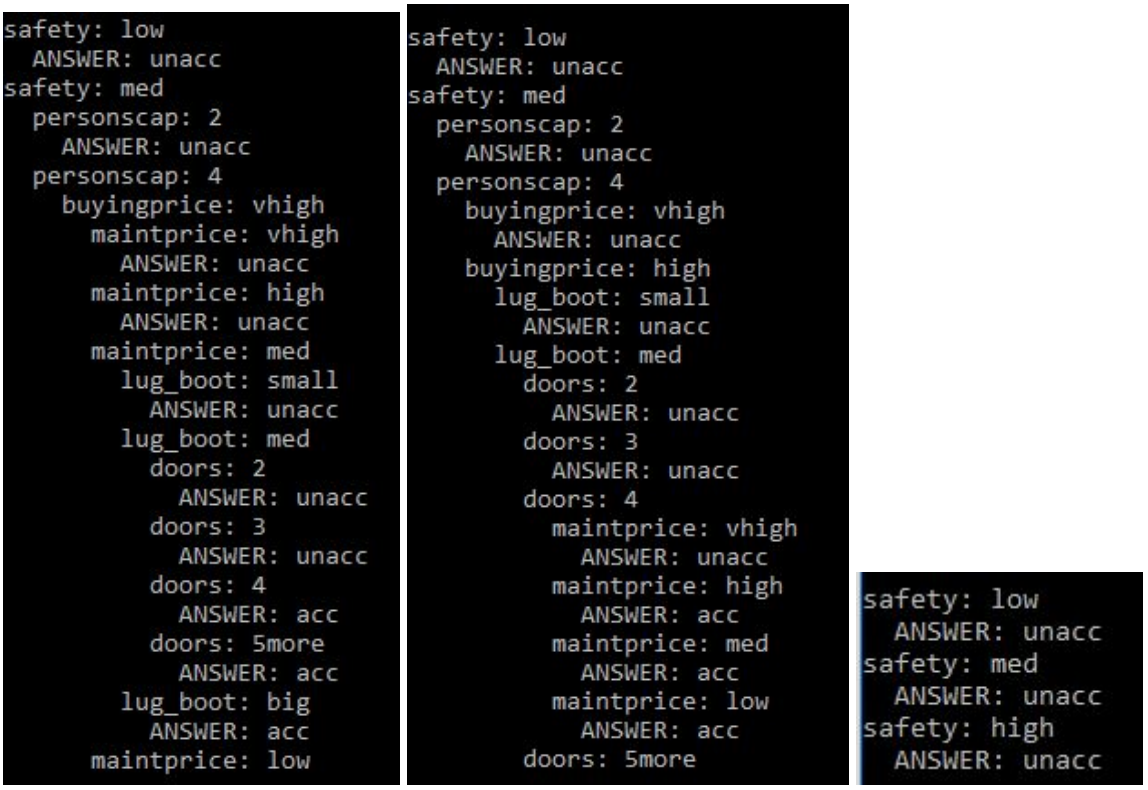
@attribute maintprice {vhigh, high, med, low}
@attribute doors {2, 3, 4, 5more}
@attribute personscap {2, 4, more}
@attribute lug_boot {small, med, big}
@attribute safety {low, med, high}
@attribute Accept {unacc, acc, good, vgood}



Weka tree for car condition data set

It is possible to observe that the output from Weka is not very useful as it is, because it is not very comprehensible. However, even if the whole tree was visible, we can see that it starts becoming very broad when the size of attributes and its values increases. If the purpose of trees is to provide with a human readable relation of the attributes, then maybe it would not be very wise to use decision trees with sets of data with lots of attributes and values.

For this dataset, we tried different values for the minimum entropy required to give a result, we started with 0, then we moved up to 0.4 and then to 0.6. With this, we could observe the effects of over and underfitting:



Our program results for car condition dataset.

The leftmost image shows part of a tree with a limit of 0 for allowed entropy. The resulting tree created around 900 nodes leading to a very accurate description of the dataset. The middle image shows part of a tree with a limit of 0.4 for allowed entropy. This tree created only about 50 nodes, resulting in a more friendly model that will more likely perform better if tested with different data sets. Lastly, the rightmost resulting tree came about with a limit of 0.6. Clearly the model is not overfitted, but this is a clear example of how trying to avoid overfitting can lead to an useless model if we are not being careful enough.

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

It is clear that decision trees are somewhat simple based on the algorithm that we have seen so far. However, they are extremely powerful when dealing with huge amounts of information where the relationships between the data are not clear on first sight. They seem to be very helpful with classification purposes where there are not too many different values an attribute can take. This can lead to a more comprehensible tree which will in turn be a lot more useful.