León Andrés Rodríguez Guzmán **A01700687**
Juan Manuel Ledesma Rangel **A01206568**
Laboratory #5
October 30th, 2018
Intelligent Systems Group #1 Tuesday and Friday 5:30pm
Rubén Stranders

15. Lab: Using Data Mining Tools WEKA

Goal: learn how to use data mining tools to rapidly test and make experiments with several machine learning algorithms **(6 hours)**

Weka is a pre-created suite that allows the creation of a decision tree in a graphical way, while the program we wrote generates text output. Although the results are the same in our program and in Weka, the last one displays the output in a more friendly manner that allows faster comprehension of the decision to be made depending on the data provided. So it can be said that an advantage of weka is the graphical representation, however a clear disadvantage is that you're limited to what the program offers, you can't change the way output is generated whereas in our program we can change the output whenever we want to that it can be in a tree like form. Another disadvantage of Weka is that you can't see what internal process is taking place to generate the output and that you need to learn how to use it, while the program we generated we already know how it processes the information (and we can display intermediate steps) and we don't need to learn how to use it. One clear difference, although not necessarily a disadvantage is that weka using J48 algorithm generates always binary trees whereas ID3 algorithm (our implementation) can have as many paths as required which becomes an advantage in certain cases..

Well, we had to take many criteria into consideration when choosing the datasets, because many of them wouldn't work for our algorithm because they had continuous data for example. So, we tried choosing algorithms that where multivariate, used for classification and the attribute type being categorical and/or integer, the number of instances and attributes was not something we considered unless we found a dataset with more than 10 attributes, because it would have been very time consuming to build the attribute section for 200 attributes for example. Other than that, we choose the the first from top to bottom that matched the already explained criteria.

The first tree constructed in weka was for a balloons dataset:

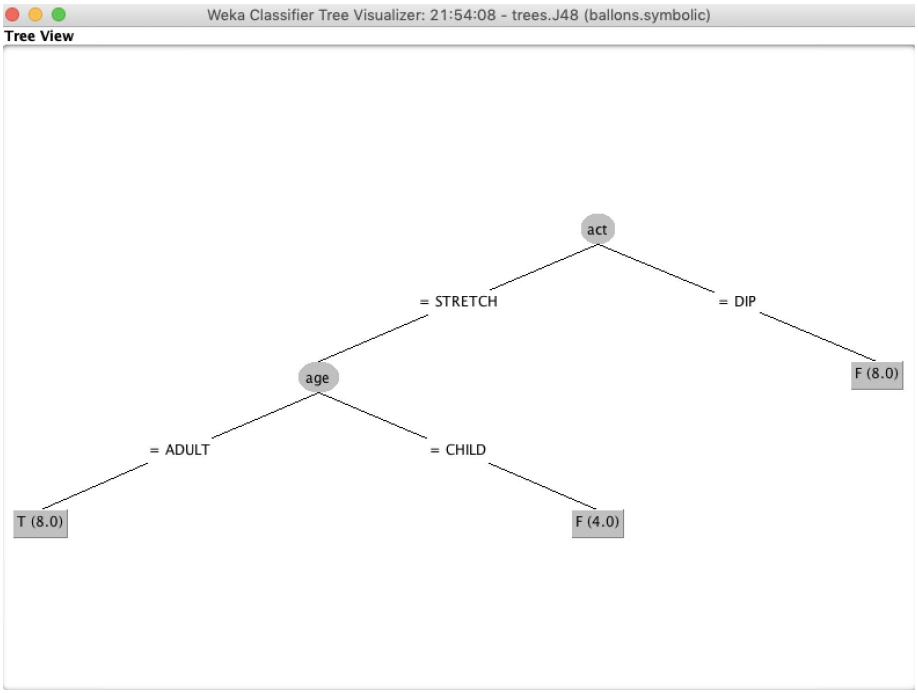

Fig 1. Balloons decision tree

And in our code we obtained:

```
JuanMa@MacBook-Pro-de-Juan DecisionTrees (master) $ ruby run.rb
@relation ballons.symbolic

@attribute color {YELLOW, PURPLE}
@attribute size {LARGE, SMALL}
@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
PURPLE,LARGE,STRETCH,ADULT,T
PURPLE,LARGE,STRETCH,ADULT,T
PURPLE,LARGE,STRETCH,CHILD,F
PURPLE,LARGE,DIP,ADULT,F
PURPLE,LARGE,DIP,CHILD,F

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

Fig 2. Balloons output in our code

We decided to test the limits of our algorithm and so we choose a dataset with more attributes and information. On Weka the tree for breast cancer looks like this:

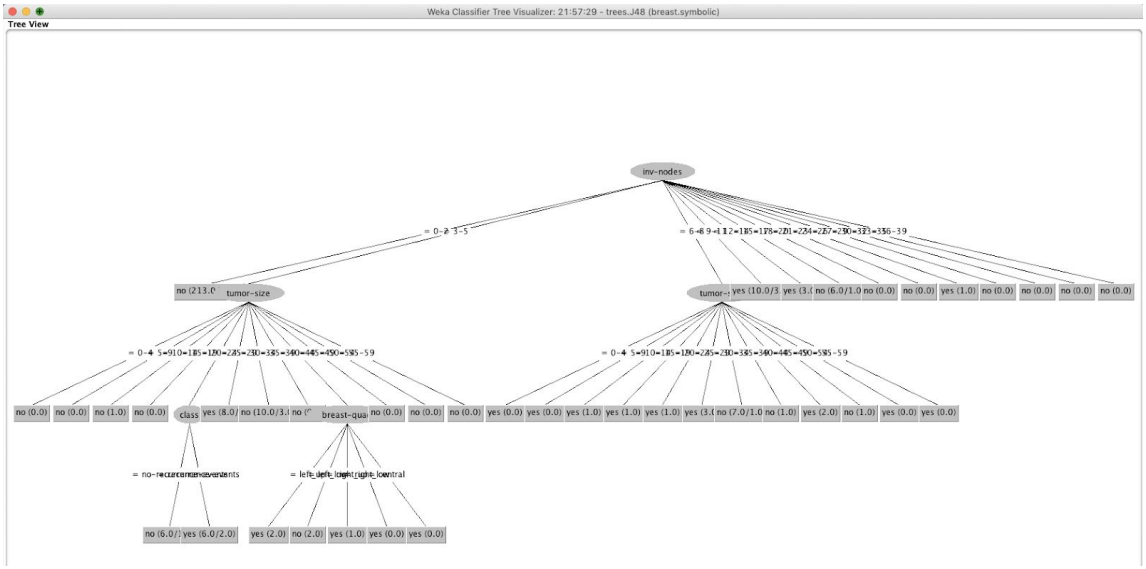

Fig 3. Breast cancer decision tree

Although not so clear an idea can be obtained of the tree structure. Now, on our code we obtained:

```
inv-nodes: 0-2
  tumor-size: 0-4
    ANSWER: no
  tumor-size: 5-9
    age: 10-19
Traceback (most recent call last):
        11: from run.rb:155:in `<main>'
        10: from run.rb:102:in `Expand'
         9: from run.rb:102:in `each'
         8: from run.rb:104:in `block in Expand'
         7: from run.rb:102:in `Expand'
         6: from run.rb:102:in `each'
         5: from run.rb:104:in `block in Expand'
         4: from run.rb:102:in `Expand'
         3: from run.rb:102:in `each'
         2: from run.rb:104:in `block in Expand'
         1: from run.rb:78:in `Expand'
run.rb:14:in `H': undefined method `[]' for nil:NilClass (NoMethodError)
```

Fig 4. Breast cancer output in our code

This is given because we have incomplete data in our dataset that is being substituted by a ?, however ID3 doesn't handle incomplete data. On Weka the tree was generated because it used J48, an extension of ID3 that can handle incomplete data. J48 algorithm does this by weights of the observed non-missing variables frequency and given those weights and the reached leaf nodes it will take a decision on splitting on the data. That means that J48 instead of

just ignoring the incomplete data or declaring a "not existing" leaf, it tries to approximate the splitting to the most proximal value that that data line could take.

As it can be appreciated, the outputs vary a little bit, when having too much attributes the Weka interface becomes difficult to read whereas in our code you can always understand the output although it's a little bit more difficult because of the list format. Besides from that and the last tested case with incomplete data, there are no more significant differences left.

Definitely! They provide a very convenient and not so difficult way to make decisions based on data. Actually, humans are all the time using decision trees when having many different alternatives. They may not notice this but the only difference between a decision tree and a human process of making a choice is that a decision tree structures data better and can process bigger amounts given computation power. Now, referring to academic or professional affairs I do think we'll be using decision trees, because they make the basis for many applications now-a-days, on chatbots for example. Decision trees are a very simple but powerful tool.

References:

Cestnik,G., Konenenko,I, & Bratko,I. (1987). *Assistant-86: A Knowledge-Elicitation Tool for Sophisticated Users*. In I.Bratko & N.Lavrac (Eds.) Progress in Machine Learning, 31-45, Sigma Press.

Clark,P. & Niblett,T. (1987). *Induction in Noisy Domains*. In Progress in Machine Learning (from the Proceedings of the 2nd European Working Session on Learning), 11-30, Bled, Yugoslavia: Sigma Press.

Michalski,R.S., Mozetic,I., Hong,J., & Lavrac,N. (1986). *The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains*. In Proceedings of the Fifth National Conference on Artificial Intelligence, 1041-1045, Philadelphia, PA: Morgan Kaufmann.

Pazzani, M. (1991). *The influence of prior knowledge on concept acquisition: Experimental and computational results*. Journal of Experimental Psychology: Learning, Memory & Cognition, 17, 3, 416-432.

Tan, M., & Eshelman, L. (1988). *Using weighted networks to represent classification knowledge in noisy domains*. Proceedings of the Fifth International Conference on Machine Learning, 121-134, Ann Arbor, MI.

Weka. (n.d). *Downloading and installing WEKA*. Retrieved from https://www.cs.waikato.ac.nz/ml/weka/downloading.html