An alternative solution for the stem rust plague in America

Adelaida Maldonado Esguerra
Universidad Eafit
Colombia
amaldonadoe@eafit.edu.co

Jesús Esteban Zapata Flórez
Universidad Eafit
Colombia
jezapataf@eafit.edu.co

Mauricio Toro
Universidad Eafit
Colombia
mtorobe@eafit.edu.co

ABSTRACT
The coffee represents a lot for many countries in central and south America. It being the major income for a lot of farmer families. But the stem rust is a threat for them because it doesn't let the coffee bean to get mature enough to sell when they fall to the ground. According to big Latin American Organization it represents a big problem for the country's economy and needs to be solved quickly.

A way to solve the problem is with decision trees like a lot of other problems that have been solved with them like noticing if a machine will stop working or if a car consumer more than what it should. And this all can be done using data structures.

1. INTRODUCTION

Coffee along history has played a very important paper in society whether is for culture, tradition or economy. Currently the coffee represents a big part of the economy for a lot of countries which most of them are located in central and south America. Because of that, a plague like stem rust destroys the crops so the country struggles economically, but even worst the farmers that base their total income in coffee lost their best revenue chance.

"Stem rust, caused by *Puccinia graminis.* Wich is a fungus that affects a lot of plantations in , especially coffee and wheat even though is more common to find it in Africa affects more the central and south America. The rust attacks the parts of the plant that are above the ground level specially the stem where it keeps the plant from reproducing by not letting it release its seed to the ground.

The history of stem rust starts since 1861 in Kenia where the first reported plague occurred. A curious fact about rust is that it was the cause why in England they drink tea instead of coffee, it was because when they suffered a rust plague in their coffee they didn't know how to exterminate it so the English decided just to drink tea and tear all the coffee plantations and plant tea.

2. PROBLEM

A lot of central American countries lay a big part of their economy on coffee production, but lately a disease that is caused by a fungus and makes that the coffee beans that are not mature enough to tear off from the plant fall to the ground has been propagating in a lot of countries which is causing that a lot of farmers lose their product.

The first time the stem rust was spotted in Central America was in 1982. But since then its propagation never had been so massive. According to BBC 10 countries in America export coffee so this plague can cause a lot of damage to the economy, also because it generates more than a million and a half of jobs in those countries. According to CropLife Latin América Organization a 30% of the jobs generated by agriculture is affected negatively by the stem rust.

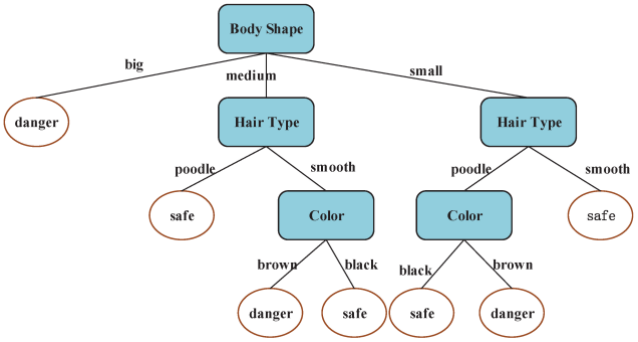
That's why a data structure solution is needed for this problem so the farmers can spot a plant that's being affect ted by the rust and heal it or pull it out the ground, and that is possible using a decision tree.

3. RELATED WORK

The stem rust problem is a detection or the assumption of a state, diagnostic kind of problem, as the following ones that are solved using algorithmic solutions like decision trees.

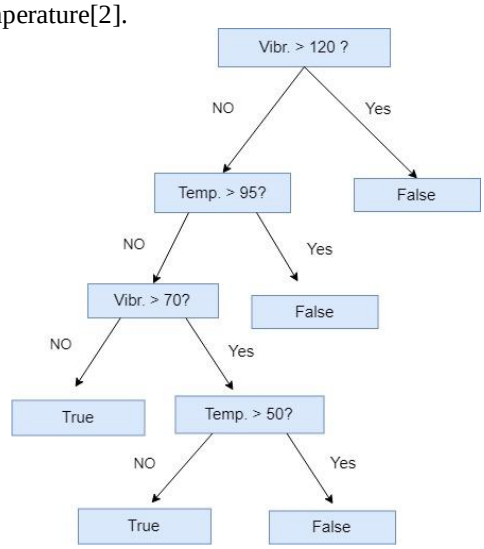
3.1 ID.3 (If a dog is dangerous or not)

This algorithmic method to create decisions trees was one of the first ones to be created an also a very basic one although very efficient. With this algorithm we assume we have a data base we can work from to take decisions from a patron by taking on consideration the gain of each category of information and with that, use the divide and conquer method to classify the mixed objects. For example, in their article Y Wang, Y Li, Y Song, X Rong, and S Zhang take a database of different types of dogs and according to their characteristics decide if they were dangerous or safe. First characteristic with more gain was the body shape because if it was big there wasn't any need to ask anything else, the dog was dangerous. But whether the dog was medium or small it would depend on the hair type to take a decision because it also had a straight forward answer but if the decision hadn't been taken yet the final factor would be the color where the tree ends[1].



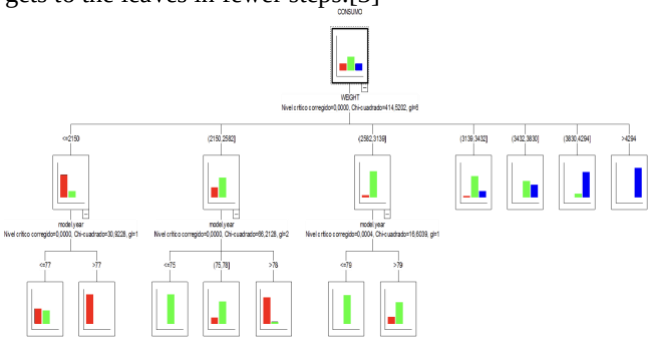
3.2 C4.5 (If a machine is going to fail or not)

The C4.5 algorithm is considered as the improvement of the id3. Specially because using C4.5 you can not only classify discrete variables but continue variables. As it predecessor this algorithm bases the criteria of which node to put on entropy. C4.5 expands the classify range to digital attribute. An other thing that this algorithm has is that it knows how to handle incomplete data points. But because of that if the data is very noisy over fitting happens so the algorithm picks up data with uncommon characteristics. Because it focuses more on gain ratio is more efficient than id3 that just focuses on information gain. What it does is that the algorithm keep dividing the data looking forward to a lower entropy so the decision can be taken easier and the C4.5 achieves it by calculating the potential data by a test to get the best attribute to branch on until end the tree. Because this algorithm can repeat attributes so the tree can become very big[3]. An example is deciding is a matching will fail (in te tree represents false), or don't (true in the tree). In the tree we can clearly see although the decision in made it had to compare the variables too many times to have an outcome, even though there were only two variables: if the machine was vibrating to much and the temperature[2].



3.3 CHAID (If a car consumes too much gas)

The CHAID algorithm was designed by Kass in the 80's. It assumes that the explicative variables are ordinals or categorical and when they are not it turns them into discrete. At the beginning it was designed for response and categorical variables, but then it also became for continue variables. The node cut supports multiple directions. CHAID considers every possible cut in every variable and selects the cut with the lowest p-value associated to a statistic contrast measurement and the hunt for the optimum cut and variable is done in two phases: the merge (where the categories fusion), and the split phase (where the cut variable is selected). An example of this algorithm being used for analyzing data and giving a positive or negative decision from it could be getting if the vehicles consume is high or low. First it takes a look on the weight of the vehicle and if it is necessary depending on the range of the weight asks for the model and the year to finally take a decision. What is good about this algorithm is that how it can take different paths asking a single question so the tree gets to the leaves in fewer steps.[3]

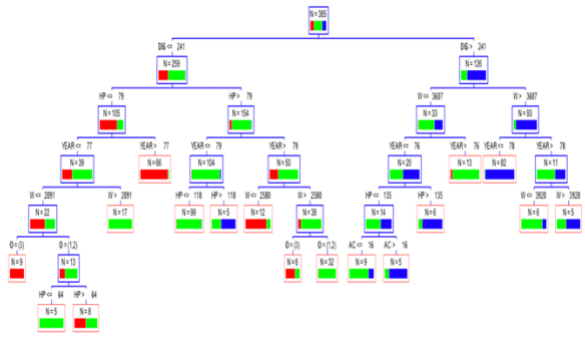


3.4 CART(If a car consumes too much gas)

This algorithm is based on statistic studies by the mathematicians from the Berkeley and Stanford universities. It works with all kind of variables, so it doesn't need to turn to discrete any continue variable. But the cut in every node is decided by a binary rule so the upcoming tree comes out deeper.

The way the CART algorithm cuts its nodes is selecting the ones that takes it to the lower entropy which makes homogeneous decreasing in the answering variable. The complexity of the tree is calculated by counting how many final nodes the tree has to then be cut and by that it means to simplify the tree. The same example of deciding if a car consumes a lot or not can also be done with CART but this time is going to have a lot of branches which means more steps to get to the leaves.

In This particular case the longest path has 6 nodes where first we check the displacement, then the horse power, the year, the weight, the acceleration and finally again the horse power. And that's only for one leave depending on which node you go a the same characteristic will be compare many times[4]



REFERENCES

1. Wang, Y., Li, Y., Song, Y., Rong, X., & Zhang, S. (2017). Improvement of ID3 algorithm based on simplified information entropy and coordination degree. *Algorithms*, *10*(4), 124.

2. Quinlan, J. R. (2014). *C4. 5: programs for machine learning*. Elsevier.

3. Wilkinson, L. (1992). Tree structured data analysis: AID, CHAID and CART. *Retrieved February*, *1*, 2008.

4. An approach for classification using simple CART algorithm in WEKA. (2017). *2017 11th International Conference on Intelligent Systems and Control (ISCO), Intelligent Systems and Control (ISCO), 2017 11th International Conference On*, 212. https://doi-org.ezproxy.eafit.edu.co/10.1109/ISCO.2017.7855983

.