COFFEE RUST DISEASE IDENTIFICATION USING DECISION TREE ALGORITHMS

Santiago Hidalgo Ocampo
Universidad Eafit
Colombia
shidalgoo1@eafit.edu.co

Ana Sofía Gutiérrez
Universidad Eafit
Colombia
asgutierrt@eafit.edu.co

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

ABSTRACT

Coffee rust is an old disease that affects coffee crops, specifically the Coffea Arabic which happens to be the main variety of agricultural exports in Colombia. The objective of this paper is to propose a way to do an early identification of it through the use of decision trees in processing data of physicochemical variables in certain crops that are related to the apparition of the disease.

Author Keywords: Decision tree, CART algorithm, classification algorithm, data structures.

Keywords of ACM: Theory of computation, data structures, design and analysis of algorithms

1. INTRODUCTION

In Colombia, coffee is one of the main agricultural exports; with an approximate production of 13.5M bags per year around 563,000 families depend on it. Several problems arise in the production process, including the rust plague, which is the main phytosanitary problem that affects coffee. Long known in coffee-growing areas of Africa, the Near East, India, Asia, and Australasia, coffee rust was only discovered in 1970 to be widespread in Brazil, the first known infected area in the Western Hemisphere. Ever since this outbreak of the disease, many countries have tried to eradicate the fungus that causes this by applying fungicide sprays and moving the crops to higher areas in which the low temperatures prevent the fungus to grow easily, but this is not a permanent solution.[4]

2. PROBLEM

Coffee rust's diagnosis is usually not made on time which leads to a lack of control of the disease, and inevitably to high production losses. There are several varieties of coffee that are more resistant to rust, however, export coffee (Caturro coffee or Coffea Arabic) is one of the most susceptible varieties to suffer from this pest. In this order of ideas, Eafit University developed a greenhouse capable of monitoring multiple physicochemical variables that are associated with the appearance of rust from a wireless sensor network, therefore, the objective of this project is to take advantage of the data that offers this system, to detect coffee rust from algorithms based on decision trees.

3. RELATED WORK

3.1 Chi-squared Automatic Interaction Detection

The CHAID algorithm, created by Kass (1980) and adapted by Magidson (1994), allows working with a categorical dependent variable (nominal or ordinal) from independent variables to establish an association or profile. [5]

It is a kind of multiple regression for categorical, discrete and discontinuous variables such as sex, socioeconomic status, religion, occupation, race, city, municipality, area, etc. in which there is a dependent variable (DV) and at least one independent variable (IV) and predicting the DV through the IV's.
The CHAID divides the population into two or more distinct groups based on categories of the "best" predictor of a dependent variable. Then divide each of these into smaller groups based on variables from other predictors. This process of continuous division ends until no more statistically significant predictors are found (or until some unemployment rule is met) and CHAID displays the final subgroups or "segments" in an easy to understand tree diagram. [2]

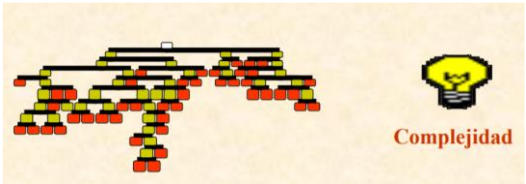

Image obtained from Data mining with decision trees: https://web.fdi.ucm.es/posgrado/conferencias/JorgeMartin-slides.pdf

3.2 Classification and Regression Trees

The CART algorithm allows to classify an specific population or define profiles under established variables of any type. It comes from the field of Statistics and was developed by mathematicians from the University of Berkeley and Stanford (Breiman, Friedman, Olshen and Stone) in the mid-80s. [1]

The process can be schematized in 4 phases: building of the tree, stopping of the tree growth process (a maximum tree is constituted in a way that over-adjusts the information contained in our database), pruning of the tree or doing it simpler and leaving only the most important nodes and, finally, selection of the optimal tree with generalization capability. [7]

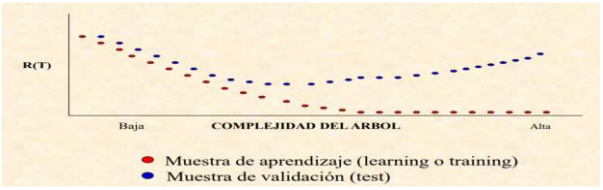

Image obtained from Data mining with decision trees: https://web.fdi.ucm.es/posgrado/conferencias/JorgeMartin-slides.pdf

3.3 ID3 (Iterative Dichotomiser 3) by Ross Quinlan

The main task of ID3 is constructing a decision tree. Nodes of the tree are labeled by attributes while arches are labeled by values of an attribute. Our assumption is that the set of examples is partitioned into at least two concepts. The attribute that is a label of the root is selected based on the maximum of information gain criterion, which means, the attribute that leads to more labeled cases of them all. The process continues until all members of the same branch belong to the same label, at this point, the tree does not need to be further partitioned. [3]

This algorithm is use in the location of branchpoints in pre-mRNA introns to provide insight into the early steps of splicing of it. A decision tree optimized by the ID3 algorithm finds the most concise and effective hierarchy of decisions to explain the process of branch-point selection by identifying the most important characteristics that determines it, being this a similar case to the one object of this study.[6]

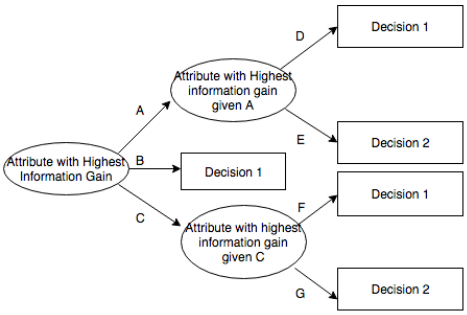

Image obtained from Potential ID3-generated decision tree. https://bit.ly/31ulBF1

3.4 C4.5 statistical classifier

Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each described by its values of attributes, and output a classifier that can accurately predict the class to which a new case belongs.

C4.5 first grows an initial tree using the divide-and-conquer algorithm depending on whether all the cases belong to the same class or there is a more prominent case. Otherwise, it chooses a single attribute with two or more out comes and makes this the root of the tree with one branch for each class and apply the same procedure recursively to each subset. Each path from the root of the tree to a leaf becomes a prototype rule whose conditions are the outcomes along the path and whose class is the label of the leaf. [8]

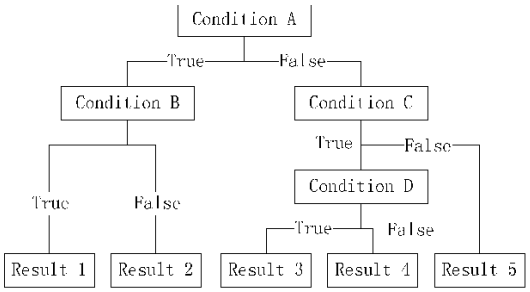

Image obtained from A Feature Selection Algorithm for C4.5: https://www.semanticscholar.org

4. Binary tree (Generated by CART algorithm)

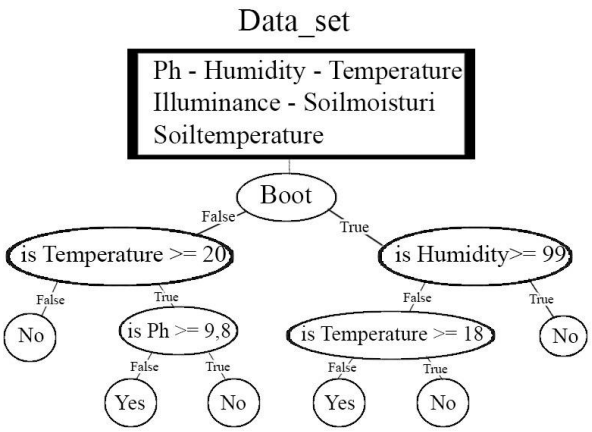

Figure 1: Tree construction example

4.1 Operations of the data structure

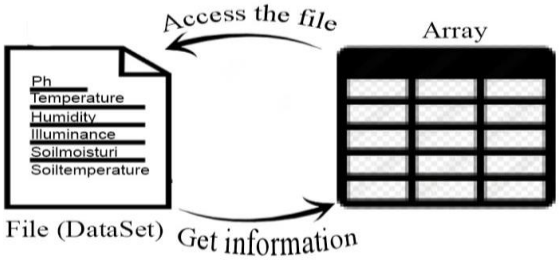

Figure 2: File reading.

4.2 Design criteria of the data structure

From the great variety of algorithms that exist to generate decision trees, the CART (Classification And Regression Trees) algorithm was chosen because it has a great

predictive capacity with respect to the other algorithms such as ID3, C4.5 or the CHAID. The most striking aspect of this algorithm is that CART selects the cut that leads to the greatest decrease in impurity. This results in homogeneous descendants in the response variable Y (in this case it is if the coffee lot has Roya or not). On the other hand, CART can work with continuous variables, which are adjusted to variables of the Data Set given.

4.3 Complexity analysis

Method	Complexity
Read data	O(n*m)
Unique values	O(n)
Is numeric?	O (1)
Find Best Split	O(n^2)
Gini	O(n)
Tree Building	O(n^2)

Table 1: Table to report complexity analysis

4.4 Execution time

	Data Set 1	Data set 2	Data Set 3	Data Set 4
File Reading	0.0028 sg	0.0029 sg	0.0038 sg	0.0028sg
Tree Building	0.3551 sg	0.5698 sg	1.5055 sg	0.7597 sg
Tree printing	0.0004 sg	0.0009 sg	0.0016 sg	0.0009 sg

Table 2: Execution time of the operations of the data structure for each data set

4.5 Memory used

	Data Set 1	Data set 2	Data Set 3	Data set 4
Memory consumption	126.2 Mb	125.7 Mb	126.7 Mb	125.9 Mb

Table 3: Memory used for each operation of the data structure and for each set data sets

5. Binary tree (Generated by CART algorithm)

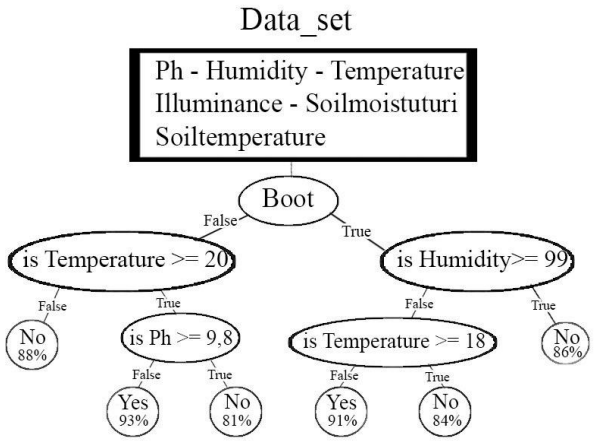

Figure 3: Tree construction example with percentages

5.1 Operations of the data structure

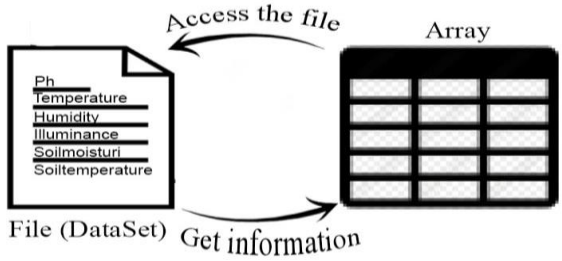

Figure 4: file reading

5.2 Design criteria of the data structure

The current data structure does not have major changes with respect to the mentioned above; therefore, the criteria of the designed data structure are the same as those mentioned in 4.2. From the great variety of algorithms that exist to generate decision trees, the CART (Classification And Regression Trees) algorithm was chosen because it has a great predictive capacity with respect to the other algorithms such as ID3, C4.5 or the CHAID. The most striking aspect of this algorithm is that CART selects the cut that leads to the greatest decrease in impurity. This results in homogeneous descendants in the response variable Y (in this case it is if the coffee lot has Roya or not). On the other hand, CART can work with continuous variables, which are adjusted to variables of the Data Set given.

5.3 Complexity analysis

Method	Complexity
Read data	O(n*m)

Unique values	O(n)
Is numeric?	O (1)
Find Best Split	O(n*m)
Gini	O(1)
Tree Building	O(2^n+m)
Is there rust?	O(2^n+m)

n: rows

m: columns

Table 4: Update table to report complexity analysis

5.4 Execution time

	Data Set 1	Data set 2	Data Set 3	Data Set 4
File Reading	0.004 sg	0.0049 sg	0.0045sg	0.0026 sg
Tree Building	0.6 sg	0.9905 sg	1.770 sg	0.7038 sg
Tree printing	0.0007 sg	0.0015 sg	0.0018 sg	0.0009 sg

Table 5: Update execution time of the operations of the data structure for each data set

5.5 Memory used

	Data Set 1	Data set 2	Data Set 3	Data set 4
Memory consumption	126.1 Mb	126.4 Mb	127.0 Mb	126.8 Mb

Table 6: Memory used for each operation of the data structure and for each set data sets

5.6 Result analysis

Below is the output of a given data set in which the presence of rust is determined:

```
Actual: yes. Predicted: {'yes': '100%'}
Actual: yes. Predicted: {'yes': '95%', 'no': '4%'}
Actual: yes. Predicted: {'yes': '88%', 'no': '11%'}
Actual: yes. Predicted: {'yes': '72%', 'no': '27%'}
Actual: no. Predicted: {'no': '85%', 'yes': '14%'}
Actual: no. Predicted: {'yes': '88%', 'no': '11%'}
Actual: no. Predicted: {'no': '75%', 'yes': '25%'}
```

¡Note the percentage of accuracy!

6. Conclusions

With this report, an algorithmic solution will be provided to determine if a given batch of coffee had rust or not given specific variables. This solution was based on the construction of a decision tree using the CART algorithm. On the other hand, the asymptotic complexity of said algorithm to determine the efficiency of the solution was revealed. It can be said that the algorithm meets the objective since most of the time it predicts correctly and with a high percentage of accuracy the state of the coffee batch. The greatest difficulty that occurred during the execution of the project was to fully understand how the CART algorithm worked, and therefore, to understand its complexity. However, it was a barrier that could be overcome through research. Finally, algorithm accuracy can be better and better by resorting to more advanced structures such as random trees and analyzing spectral images.

6.1 Future work

To advance to a much more accurate solution, more complex data structures such as a random forest must be used, thus reaching much longer percentages of accuracy. The advantage of implementing a random forest is that it is based on the construction of several decision trees, so the current structure would be the basis for a more precise solution. However, if you want to reach a much higher accuracy, you must implement algorithmic models that analyze multispectral images, so the data obtained will be much more accurate.

ACKNOWLEDGEMENTS

We are deeply grateful to Isabela Piedrahita Velez, teacher assistant of the course "Structures and Algorithms 1" Eafit for his constructive comments about algorithm optimization.

REFERENCES

1. Arevalillo, J.M. Data mining con árboles de decisión. Retrieved from Universidad Nacional Educación a Distancia, Madrid: https://bit.ly/2M9ZL61

2. Bello, P., León, D. Técnica del Árbol CHAID. 2013. retrieved from: https://bit.ly/2Kt0e0X **3.** Jerzy, W. Selected Algorithms of Machine Learning from Examples. *Fundamenta Informaticae* 18 (1993), 193–207. Retrieved from Department of Computer Science, University of Kansas: https://bit.ly/2TtEuoH **4.**Petruzzello, M. Coffee rust: disease, 2017. Retrieved from Encyclopaedia Britannica: https://bit.ly/2GVrsvd

5.Sanz, E., Ponce de León, Ana. Claves en la aplicación del algoritmo CHAID. un estudio del ocio físico deportivo universitario, 2010. *Psicología del Deporte*. 19 (2). 319-333. Retrieved from Universitat de les Illes Balears: https://bit.ly/2yUOopO **6**.Taggart, A., DeSimone, A.M., Shih, J.S., Filloux, M.E., Fairbrother, W. Nat Struct Mol Biol. *US National Library of Medicine*, 2012. 19(7). 719–721. Doi: 10.1038/nsmb.2327

7.Trujillano, J., Sarria, A., Esquerda, A., Badia, M., Palma,

M., March, J. Approach to the methodology of classification and regression trees, 2008. Retrieved from: https://bit.ly/2H2zu5z
8.Xindong, W. et al. Top 10 algorithms in data mining. *Know Inf Syst*, 2008. 14. 1–37. Doi: 10.1007/s10115-007-0114-2