COFFEE AND RUST: DETECTION AND PREVENTION FOR IMPROVING EXPORTATION QUALITY

Isabella Quintero
Universidad EAFIT
Colombia
iquinterov@eafit.edu.co

Sofia Vega
Universidad EAFIT
Colombia
svegal@eafit.edu.co

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

ABSTRACT

Coffee is a very important product around the world, it exists a huge market around it, and the economy of countries such as Brazil, Vietnam and Colombia, is based on its production and commercialization.

Along the years the countries that produce coffee have experienced a lot of problems, the most relevant is Rust, it is caused by the fungus Hemileia vastatrix, this plague has threatened Latin-America's and Africa's crops, it infects the leaves on the coffee plant, which makes it not suitable for human consumption, and consequently it cannot be exported anymore, this causes very large loses of money and time. What makes this problem so big and uncontrollable is that when farmers can detect it, is too late to handle.

That is why, it becomes necessary to be able to monitor in real time the things that make rust appears in crops; in order to do this we came up with the idea of organizing all the data in binary trees, so the evaluation of the values would become a lot easier, and the predictions, based on these values, would be more effective, with the accuracy that we obtain we were able to predict if there is going to be rust in specific conditions; furthermore we can conclude that with the given data is not very likely to have a high accuracy, so it would be better to have more characteristics of the coffee leaf.

Keywords
Decision tree, physical-chemical factors, early detection, prevention.

ACM CLASSIFICATION Keywords

CCS → Applied computing → Computers in other domains → Agriculture

CCS → Applied computing → Life and medical sciences → Computational biology → Biological networks

1. INTRODUCTION

Nowadays, it is very common for almost everybody to drink a cup of coffee regularly, what we do not always notice, is all that happens in order to produce this simple cup. Coffee producers around the world have confronted different problem related to the coffee plantation, one of them more harmful than the others: Rust.

"The current coffee rust outbreak dates back to 2012. It is a continuing problem, disrupting coffee-growing activities throughout Central America, where more than 1.3 million people depend on the cultivation of this crop. The damage caused by this infestation is compounding the effects of the fluctuation in international coffee prices, particularly since 2013, and the drought conditions in 2015 affecting the quality and quantity of coffee production. One of the main effects of the low coffee yields has been a reduction in incomes and employment opportunities, particularly in the case of small coffee growers. This is the hardest hit population due to its lack of economic means, preventing them from engaging in good crop management practices such as applying fertilizers and fungicides, among others, and their lack of agricultural education for the proper management of coffee plantations." [1]

For that reason, the purpose of this project is to create a solution by being able to analyze each and every variable involved in the infection and spread of rust in coffee plants, so the problem can be detected in real time, and, then, it can be controlled.

2. PROBLEM

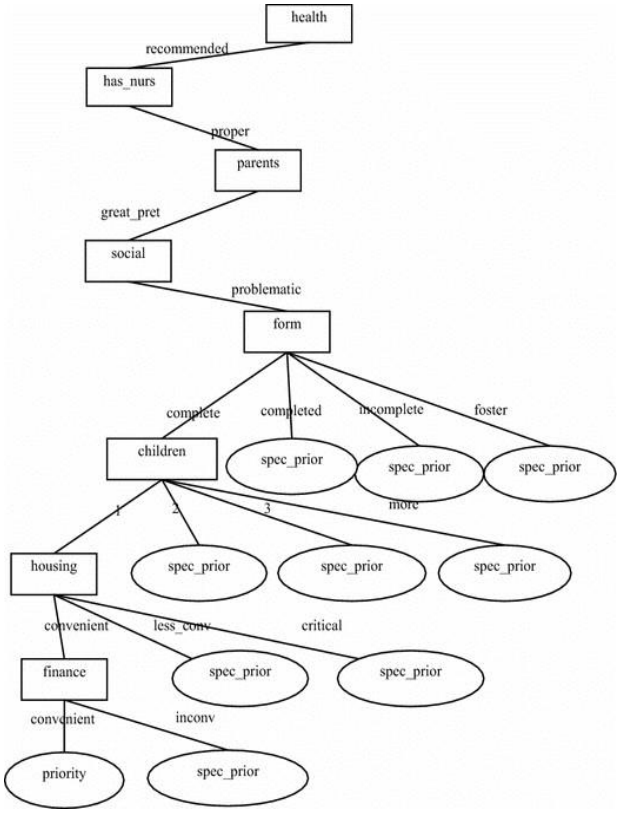
In Colombia, Coffee is the main exportation product, in particular, Caturra variety (coffea arabica) is the most important part of this economy, but it has a very low resistance to rust plague, the prime problem that affects coffee around the world. Besides, the detection of this plague in crops takes a lot of time, therefore, it is very hard to control, which leads to large loses.

3. RELATED WORK

3.1 ID3 Algorithm

This algorithm constructs a decision tree by a set of examples, is used for classification of future instances. Each example has various attributes that belong to a class, the tree leaf nodes contain the name of the class, while non-leaf nodes are the decision nodes where each one of them, corresponds to a possible attribute value. Every node of decision is a proof of the attribute with other tree that begins from it, the algorithm ID3 use the "information gain" to decide which aspect goes in each node of decision, with this we can define how well an attribute divide de examples of training in every class; it uses a concept named "Entropy" that corresponds to the quantity of an attribute´s uncertainty

and it´s formula equals to $\sum -px\ log2\ px$, where x is the set of classes in S and px is the proportion of S that belongs to the class x.



Branch of decision tree based on ID3 algorithm [5]

3.2 C4.5 Algorithm

This algorithm is the successor of ID3; the difference is that C4.5 converts the trained trees into sets of if-then rules. Each one is evaluated for their accuracy looking for established the order in which they should be applied, when this thing happens is called "Pruning" and means removing a rule´s precondition if the accuracy of the rule improves without it.

4.2 Design criteria of the data structure

To choose this data structure we based mainly in the amount of variables we were taking into consideration, therefore, the tree set begins with label, this variable practically breaks in half all of our data, then, we use all the others variables, so we can obtain an improved classification, and consequently analyze more carefully in which cases rust can be detected in coffee plantations, leading us to determine the ideal environment conditions where coffee will be free from rust, and helping us prevent this infection.

3.3 C5.0 Algorithm

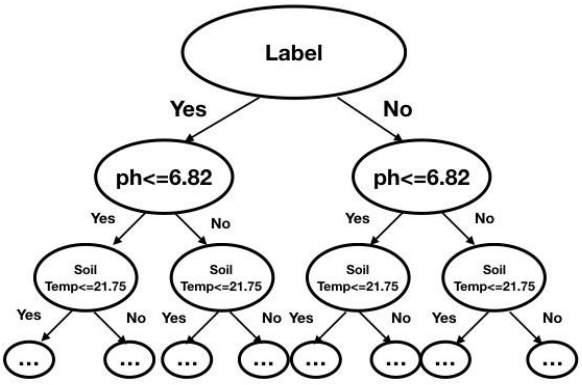
This algorithm constructs the trees based on a set of data of the optimized training under the criteria of information gain and corresponds to an evolution of its last version, the algorithm C4.5; the biggest advantages of this version are related with the efficiency in the tree´s construction time, the memory usage and the obtaining of shorter trees than in C45 algorithm, with the same predictive capacity.

Additionally, has the option of consider some attributes in order to focus the construction of the tree and being able to use a penalized learning can be used in which it is possible to assign a value to the possible outcomes.

3.4 CART Algorithm

This algorithm is a binary decision tree that is constructed by splitting a node into two little node recursively, beginning with the principal node, the root, that contains the whole learning sample; each split depends on the value of only one predictor variable, if X is a nominal categorical variable of I categories, there are (2I-1)-1 possible splits for this predictor; on the other hand, if X is an ordinal categorical or continuous variable with K different values, there are K-1 splits on X.

4. Designed Data Structure: Tree Set



4.3 Complexity analysis

Method	Complexity
Info()	O(n)
Infox	O(n)
Gain()	O(1)

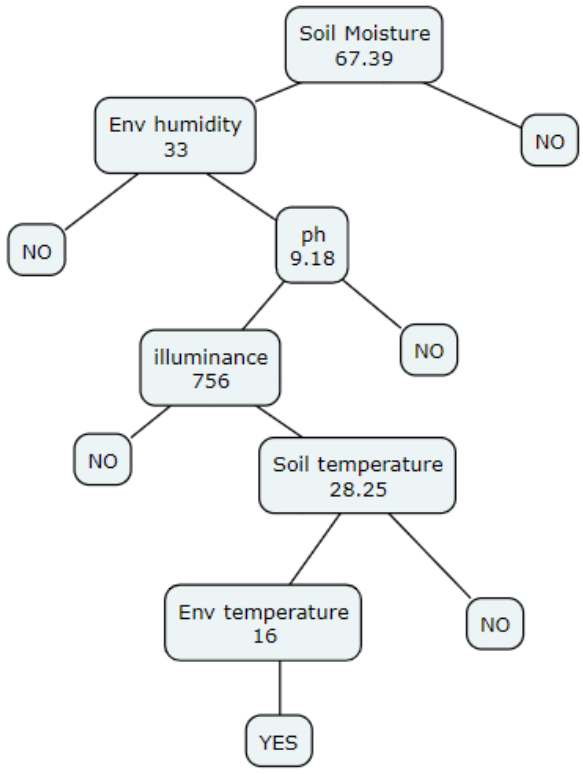
5. Decision tree (Last Implementation)

Before organizing the data onto a decision tree, we upload it to the program in a matrix, in which rows are the instances of all the data, and columns are each characteristic given.

ph	soil_temperature	soil_moisture	illuminance	env_temperature	env_humidity	label
6.44	21.0	65.22	1431.0	19.0	99.0	yes
6.23	27.0	19.2	1204.0	36.0	42.0	yes
7.53	24.5	48.55	3303.0	26.0	87.0	yes
7.33	24.5	32.97	5437.0	26.0	79.0	yes
7.07	22.25	49.28	3270.0	25.0	99.0	yes
6.9	27.0	50.36	2154.0	30.0	43.0	yes
6.49	20.75	52.9	1429.0	19.0	99.0	yes
6.52	21.5	52.9	5005.0	28.0	92.0	yes
6.51	25.5	21.01	4872.0	35.0	51.0	yes
6.54	26.25	23.55	1275.0	27.0	85.0	yes

Matrix of the Data.

Then the data is classified in a decision tree, in the following way



Decision Tree given by the program

5.1 Operations of the data structure

Our program focuses on making the best classification of the data, therefore, the operations presented here are going to be explained based on statistic concepts.

5.1.1 Calculate Entropy

Entropy tells us how homogeneous is the information/data that we have. This value goes from 0 to : if it is 0, it means the data is completely homogeneous or, in other words, the same in all cases; on the other hand, if this value is 1, it means the data is perfectly divide/classified, or each case has the same amount of values. This component is calculated in the following way

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

Where p_j is the probability of the case, which means, number of specified cases between total of cases.

5.1.2 Information Gain

This value gives us the best split for building the tree, it will be better when its bigger, because then the information will have the capacity to give a more accurate prediction due to the fact of a better exploitation of the data.

Using both of these concepts, the program looks for the best split creating subsets below the nodes of the tree, and leading it to the minimum leaves possible, that way we can obtain the best accuracy, which is the goal here, so we can predict correctly and protect the coffee crops.

5.2 Design criteria of the data structure

We needed a data structure that could have the ability to transform and organize the amount of data given by the observations Therefore we chose a decision tree, it guarantees a right split of the data, which also gives us an accurate prediction to label properly the coffee crops.

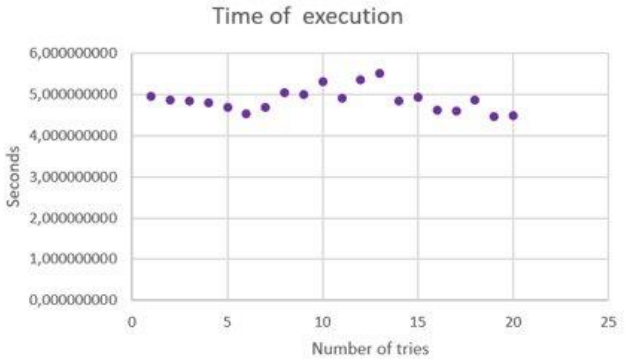
5.3 Complexity analysis

Operation	Complexity
calculateEntropy	O(n*m)
findDecision	$O(n_2 * m_3)$
BuildDecisionTree	$O(n_3 * m_3)$

Table to report complexity analysis

In the previous analysis n is the number of columns or features of the data, and m is the number of values for each column.

5.4 Execution time



Measure the execution time and memory used 100 for each data set and for each operation of the data structure. Report the average values.

5.5 Memory used

Memory consumption	140,9MB

Memory occupied by the program

5.6 Result analysis

Through the use of this algorithm and data structure we obtained an average accuracy of 89% using random forest and ID3 algorithm; besides we also obtained an accuracy of 85.6% in prediction using only C4.5 algorithm, that is because this last algorithm is more effective and precise. These results show us a direct relation between environment conditions and appearance of rust in coffee leaves, which can lead us to a more efficient prevention of the disease.

6. CONCLUSIONS

In conclusion, coffee rust is something that affects huge areas of the world's economy and if we find the way to connect the technology with the nature, it can help a lot to increase the gains of the countries that export coffee to all around the world, with more characteristics to create the decision trees it could have better accuracy and finally we could predict with a lot of certainty in which conditions the coffee rust grows and also we can control the dispersion of it.

With this project, we were able to properly classify information and learned how to read and analyze it. On the other hand, we also could do predictions highly accurate thanks to the use of decision trees and C4.5 algorithm.

6.1 Future work

This project does not accomplish efficiently to predict either there will be or not rust in a coffee crop, due to the fact that it does not take into consideration many other variables that also take a part in this problem, besides, if we also take those variables, we will increase the time and memory consumption of the program, making it considerably less efficient. Therefore, we will need a better algorithm, or additional features to the existent program that make easier the prediction and do not affect the efficiency of it.

ACKNOWLEDGEMENTS

We thank for support with coding and development to Mateo Bonnett García, student of Mathematical Engineering in EAFIT.

REFERENCES

1. The impact of the coffee rust outbreak on the coffee sector in central America, 2016. Retrieved August 09, 2019 from Fews net: https://fews.net/sites/default/files/documents/reports/CENTRAL%20AMERICA %20-%20Special%20Report%20-%20Coffee%20Sector%20-%202016.pdf
2. Dupouy, C. Aplicación de árboles de decisión para la estimación del escenario económico y la estimación de movimiento la tasa interés en Chile, Universidad de Chile, 2014, retrieved July 14. 20- 23.
3. CART Algorithms, 2019, 2013. Retrieved August 08, from IBM knowledge centerhttps://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.statistics.algorithms/alg_tree-cart.html
4. Tree algorithms: ID3,C4.5, C5.0 and Cart, 2019. Retrieved August 08, 2019, from Medium Corporation US:https://medium.com/datadriveninvestor/tree-algorithms-id3-c4- 5-c5-0and-cart-413387342164
5. https://www.researchgate.net/figure/Branch-of-decisiontree-based-on-ID3- algorithm_fig3_323401116
6. serengil(2019). Lightweight Decision Trees Framework supporting Gradient Boosting.