ROYA IN COFFEE PLANTS, A PARASITE TO BE ELIMINATED WITH DECISION TREES

Daniel Alejandro Cifuentes Londoño
Universidad Eafit
Colombia
dacifuentl@eafit.edu.co

Cristian Alexis Giraldo Agudelo
Universidad Eafit
Colombia
cagiraldoa@eafit.edu.co

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

ABSTRACT

Traditionally, different types of coffee have been planted in Colombia and for a long time, it has been a fundamental basis for the economy of this state, significantly projecting its image abroad and guaranteeing primary profits not only to the government but also to the growers around the country.

Competitiveness in the market with countries such as Brazil and Vietnam immediately involves the quality of the crop, since that is why the acquisition abroad of the said product depends. This quality has been affected by a parasite called rust, due to the negative effect, it produces on large plantations, thus presenting a sharp decrease in the rate of production and therefore in the general economy of the country.

For these types of concerns, the task is to find a functional system that allows the early discovery of this pest, using a set of wireless sensor networks that collect physicochemical data. With this information, early detection of the rust in the crops is obtained using decision trees, so it is controlled gradually and does not affect any system that is related to this problem.

Keywords

- Reading files
- Data structure
- Search tree
- Recursive tour
- Data storage
- Operations
- Notation O
- Execution times
- Memory
- Complexity

ACM classification keywords

Theory of computation → Computational complexity and cryptography → Problems, reduction and completeness

Theory of computation → Data structures design and

analysis → Sorting and searching.

Software and its engineering → Software organization and

properties → Operating systems → File systems

management

Theory of computation → Design and analysis of algorithms → Graph algorithms analysis → Shortest paths.

1. INTRODUCTION

Coffee has always been an element of export appreciated worldwide since the 19th century. This is initially marketed in Europe and produced in most of Asia and Africa, where it had a great reception by this population, however in the middle of 1869 when its production was rising strongly, a fungus appears which they called Hemileia vastatrix, which It

gradually affected production. At the time there was no effective cure for this pest and all the crops were moved mainly to America. Everything was going well until in 1970, this plague appears again in countries such as Colombia and Brazil where the productions were highly considered, since then the climate change, the constant rains and the tropic aspect of the region have not found an effective solution until the present.

Due to this, it is proposed to present in this document some data structures, where the importance of the use of decision trees is provided to allow to generate an assertive resolution through the use of physicochemical aspects that give certified solutions to the millions of growers and can establish a correct data analysis with balances and all types of mishaps, the complexities that this planning presents will be used like formal conclusions which allow a better perception of the problem and give control over the best suitable solution to settle positively in all systems.

2. PROBLEM

The presence of rust in different crops and productions has become an infallible phenomenon and solutions with pesticides and all kinds of chemicals have not given good results.

Therefore, it seeks to prevent the presence of this fungus through algorithms that implement sequences such as decision trees that take different types of information, precisely those physicochemical details that are detailed of the surface and the environment.

Thus giving a standard control over all crops, which in turn drives the general economy of the state and does not allow me that thousands of plantations suffer significant losses and even to develop these methods in other applications of different types of crops thus benefiting each system that involves a similar problem.

3. TRABAJOS RELACIONADOS

3.1 ID3

[1] In 1979, J. Ross Quilan created the ID3 algorithm which uses artificial intelligence that encompasses the search for a thesis or rules through a set of examples.

It is used to generate a Decision Tree from a dataset. ID3 is also considered as a precursor to C4.5, as well as many other decision tree algorithms. A big advantage of using the ID3 algorithm is that it builds the fastest but also the shortest trees; this makes it more understandable the problem.

Its use is included in the search for hypotheses or rules in it, given a set of examples. The set of examples must be made up of a series of tuples of values, each of them called attributes, in which one of them, (the attribute to be classified) is the objective, which is of binary type (positive or negative, yes or no, valid or invalid). In this way, the algorithm tries to obtain the hypotheses that classify before new instances, if this example is going to be positive or negative. ID3 performs this work by building a decision tree. The elements are:

• Nodes: Which will contain attributes.

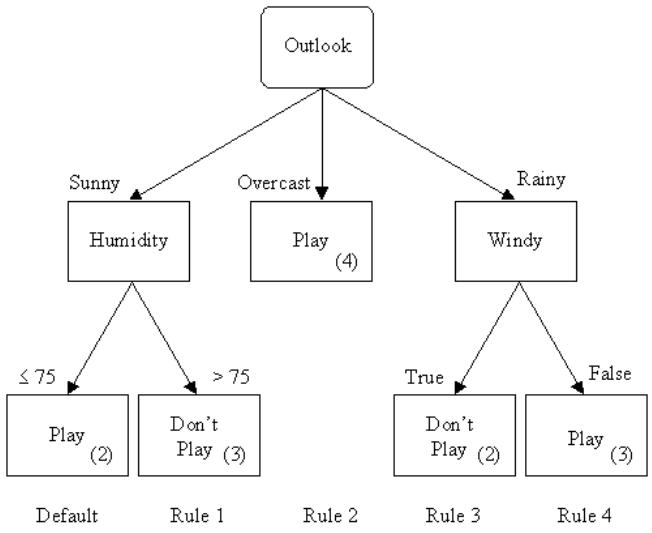
• Arcs: Which contains possible values of the parent node.

• Leaves: Nodes that classify the example as positive or negative

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

3.2 C4.5

Is the successor to ID3 and removed the restriction that features must be categorical by dynamically defining a discrete attribute (based on numerical variables) that partitions the continuous attribute value into a discrete set of intervals. C4.5 converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. This accuracy of each rule is then evaluated to determine the order in which they should be applied. Pruning is done by removing a rule's precondition if the accuracy of the rule improves without it.



[2] Characteristics of Algorithm C4.5

• It allows working with continuous values for the attributes, separating the possible results

in 2 branches Ai <= N and Ai> N.

• Trees are less leafy since each leaf covers the distribution of classes, not a

particular class.

• Use the "divide and conquer" method to generate the initial decision tree from a

training data set.

• It is based on the use of the gain ratio criterion, defined as

I (Xi, C) / H (Xi). In this way, it is possible to avoid that the variables with a greater number of

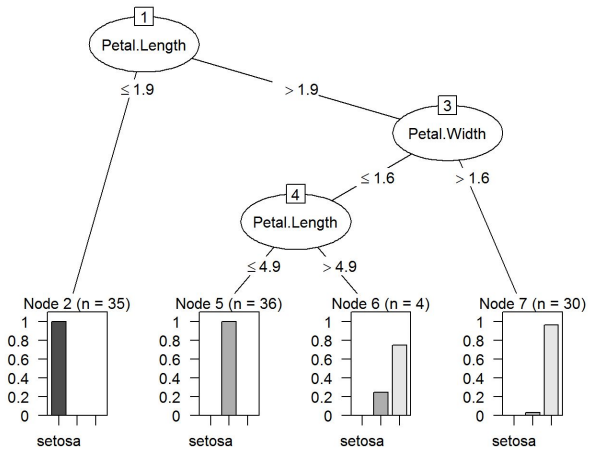
Possible values benefit from the selection.

• It is Recursive.

3.3 C5

[3] It is also the successor of C4.5, this is developed by Ross Quinlan, within its operation, divides the information and through the intervals, it can work without having all the data collection in its favor because it predicts its successions and in turn generates greater speed in decision making.

C5.0 can generate two types of models. A decision tree and a simple description of the divisions that have been found in the algorithm. The different terminal nodes (or "leaf") describe a subset of training data, and each of the cases included in the training data belongs exactly to a terminal node of the tree. In other words, it is possible to make exactly one prediction for each specific data record present in a decision tree.

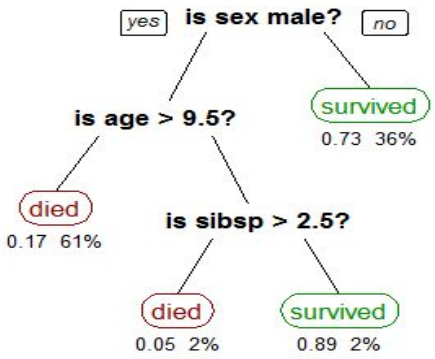


3.4 CART Algorithm

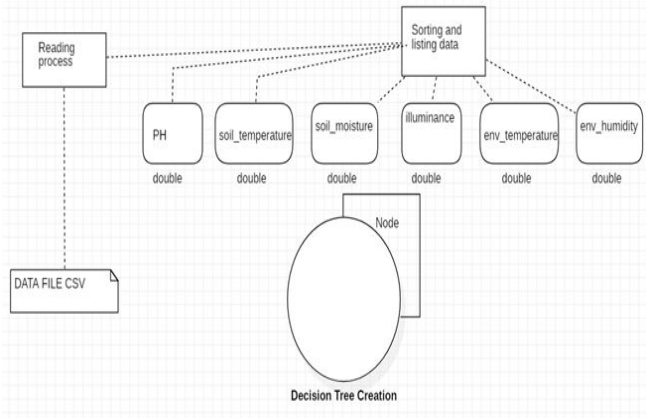
[4] The CART algorithm is based on Classification and Regression Trees by Breiman et al (1984). A CART tree is a binary decision tree that is constructed by splitting a node into two child nodes repeatedly, beginning with the root node that contains the whole learning sample

Its structure is very similar to other decision trees' but its difference with the rest is in the way the tree is built.

The CART algorithm is based, instead of Entropy and information gain values, in the Gini index function, which simplifies the way to define the pureness of the nodes and leaves.



4. Data structure



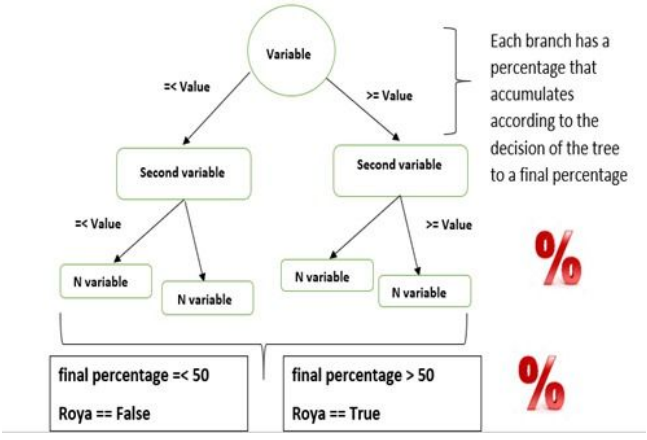
Graph 1: Implementation of a tree-like structure.

First, the data from a CSV file is read, then sorted and converted to a double type.

A structure is created where each node will have a direct relationship with the other nodes. They are divided into the parent node which is the main value from which we will begin to evaluate the variable and thus the child nodes will be evaluated sequentially depending on the previous node.

4.1 Data structure operations

ph	soil_temperature	soil_moisture	illuminance	env_temperature	env_humidity	label
6.98	25	58	4320	36	53	no



Graph 2: Rust detection process

This operation determines the value of the first variable. Depending on the value, it stores a percentage in an external variable, in addition, it sequentially passes to another node where another variable evaluates that numerical value and according to the range in which that numerical value is found, represents a percentage that is added to a final percentage.

In the end, if the accumulation of the percentage is less than or equal to 50 there is no rust, but if it is greater than 50 the plant has rust.

4.2 Design criteria of the data structure

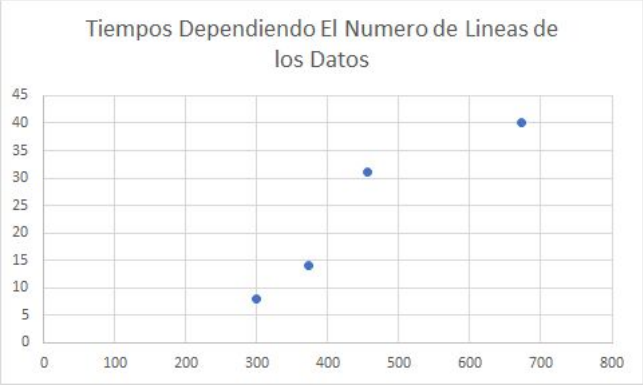
This data structure is interpreted in an appropriate way for each person who wishes to analyze these operations, making it adaptable to each change that has to be made.

Apart from this, this structure is very useful for what you want to achieve, the complexity of its methods allow the execution time to be fast. This can be evidenced in the time table of this document. These times do not exceed a high average of seconds, which means that its efficiency is a strength in this structure and adapts properly to any other implementation, that is, not only to use it to detect rust but also to some other factor that influences society.

4.3 Complexity Analysis

Metodo	Complejidad
Crear LinkedList	O(1)
Crear arboles	O(log n)

4.4 Execution Times



4.5 Memory

	Caso 1	Caso 2	Caso 3	Caso 4
Consumo Memoria	**4 MB**	**7 MB**	**8 MB**	**10MB**

REFERENCES

[1] Bsea, Dankz, Farisori, Jesuja, LordT, Paintman, Pinar, Superzerocool, Tano4595, Varano. ~ Algoritmo ID3 ~ . 2014. Retrieved August 11 2019. From https://bit.ly/2MV9exD

[2] Espino, Tijerina, Cedano, de la Fuente , Pérez, Chiñas. ~ ALGORITMO C4.5 ~ . November 2005. Retrieved August 11 2019. From http://bit.ly/2yTgnq2

[3] IBM. ~ NODO C5.0 ~ . . Retrieved August 11 2019. From https://ibm.co/2Z3Kaak

[4] Amir Ali. ~ Decision Tree (CART) Algorithm in Machine Learning ~ . July 2018. Retrieved August 11 2019. From https://bit.ly/2ZWCOGO