COFFEE RUST DETECTION FOR CATURRA VARIETY USING DECISION TREES

Paulina Pérez Garcés
Universidad EAFIT
Colombia
pperezg@eafit.edu.co

Santiago Cartagena Agudelo
Universidad EAFIT
Colombia
scartagena@eafit.edu.co

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

ABSTRACT
Coffee is one of the main plant-based goods consumed today, as well as the second most traded commodity in modern economy after oil. Its production takes place mostly in developing countries such as Colombia, where a malfunction during the process could affect thousands of coffee farmers. Coffee rust does just that, reduce the quality and quantity of grains devastating entire productions yearly. [1]

One of the most effective ways to prevent this disease from spreading is an early detection, which makes effective the extraction and replacement of withered crops. Using wireless sensor networks, it is possible to obtain data from each plant's conditions; which allows us to determinate whether a specific plant is in the early stages of infection or perfectly healthy.

1. INTRODUCTION

Coffee has been one of the most popular drinks since the 1800s, when all of eastern British colonies began growing it in large quantities. Around 1879 coffee farmer in Sri Lanka (formerly called Ceylon) noticed that a new fungus called Hemileia vastatrix was affecting their plantations, but an efficient cure was never found.

The losses caused by this disease made it unsustainable to keep growing coffee in those regions; therefore, plantations of a single healthy strain -Arabica coffee- were moved to the Americas, which quickly became the biggest productors of this good. The crops there were constantly checked, and remained rust-free until 1970, when spores from the fungus were found in Bahia, Brazil. [2]

Since arabica coffee was the most planted strain in America, and due heavy winds and rainy seasons, rust spread around the continent in a matter of years. Countries where the plant had become one of the main tradable products began suffering substantial economic losses.

2. PROBLEM

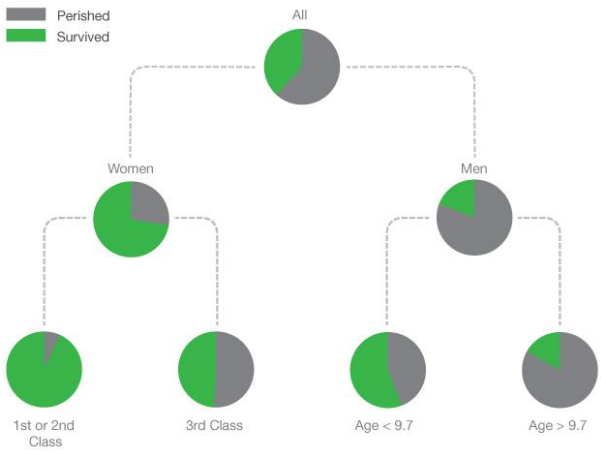
The intent of this project is to create an algorithm capable of determinate whether a coffee bush is vulnerable and suffering from rust or not. These results will be based on data such as soil temperature, weather temperature, soil's PH, illuminance and others; all taken from wireless sensor networks placed in a greenhouse in University EAFIT in Medellín, Colombia.

Said algorithm will be making decisions using methods known as decision trees. This approach uses statistics from the previously classified crops to identify new affected plants, in more technical words, it is a supervised learning method used for classification and regression. [3]

3. RELATED WORK

3.1 CART

CART is an alternative decision tree algorithm. It can handle both classification and regression tasks. This algorithm uses a metric called "Gini Index"; whose objective is to create decision points for classification tasks. CART is similar to C4.5 but there is a notable difference between them: CART constructs the binary trees based on numerical splitting, whereas C4.5 include the step of rule sets. The CART algorithm is structured as a sequence of questions, the answers to which determine what the next question, if any, should be. The result of these questions is a tree like structure where the ends are terminal nodes at which point there are no more questions. [4]



[5]

3.2 CHAID

This algorithm, short for Chi-squared Automatic Interaction Detector is one of the oldest (1980) and most popular decision trees. It is a simple set of rules that lets you add more than two tree branches on a single node. The name, as well as some of its principles, comes from the Chi-Squared Statistic that expresses the relationship between two or more non-numerical variables. This technique can easily analyze

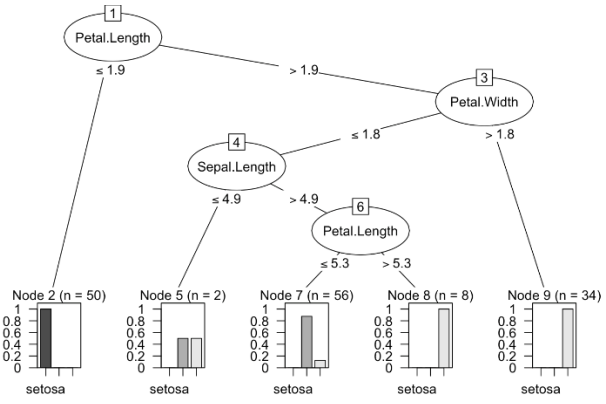
large amounts of data, but that would also create big amounts of divisions inside the tree. [6]

3.3 ID3

One of the most popular algorithms is the ID3 algorithm, also known as the Iterative Dichotomiser 3. 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 ID3 algorithm is that it builds the fastest but also the shortest trees; this makes more understandable the problem. Just as all the algorithms used for building decision trees it has a couple of disadvantages, the principal one being that if you want to implement ID3 algorithm is expensive to train it, meaning you need a lot of data to do so. [7]

3.4 C5.0

C5.0 algorithm is the successor of C4.5, developed by Ross Quinlan in 1993. It splits decisions from the one that gives more information to the ones that can easily be removed or ignored. When the test data comes through, it predicts a single category to which it belongs.; these classes into which the data fall can be discrete attributes divided into a set of intervals. C5.0 trees can be very useful when data is missing from datasets, because it works relatively well without it. The main differences with its processor are the speed, memory usage and size of the trees. [3][8]



[9]

REFERENCES

1. Goldschein, E. 11 Incredible Facts About the Global Coffee Industry. *Business Insider*. 2011. https://www.businessinsider.com/facts-about-the-coffee-industry-2011-11?op=1#according-to-the-world-wildlife-fund-of-the-50-countries-with-the-highest-deforestation-rates-from-1990-to-1995-37-were-coffee-producers-11

2. Arneson, P.A. Coffee rust. *The Plant Health Instructor*. 2000. https://www.apsnet.org/edcenter/disandpath/fungalbasidio/pdlessons/Pages/CoffeeRust.aspx

3. Scikit-learn Documentation. Decision Trees. https://scikit-learn.org/stable/modules/tree.html

4. Brownlee, J. Classification And Regression Trees for Machine Learning. *Machine Learning Mastery*. https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

5. The Fact Machine. http://www.thefactmachine.com/wp-content/uploads/2014/05/CART-TreeMap6Node.jpg

6. Statsoft. Popular Decision Tree: CHAID Analysis, Automatic Interaction Detection. http://www.statsoft.com/Textbook/CHAID-Analysis

7. Howe, B. Building Trees: ID3 Algorithm. Coursera. https://es.coursera.org/lecture/predictive-analytics/building-trees-id3-algorithm-7PNNn

8. IBM. C5.0 Node. https://www.ibm.com/support/knowledgecenter/en/SS3RA7_15.0.0/com.ibm.spss.modeler.help/c50node_general.htm

9. Kuhn, M; Ross, Q. Plot a decision tree. https://topepo.github.io/C5.0/reference/plot.C5.0.html