Wireless sensors for detecting rust in caturra coffee: Data structures for the prediction of infected crops.

Juan Pablo Ossa Zapata
Eafit University
Colombia
jpossaz@eafit.edu.co

Mauricio Jaramillo Uparela
Eafit University
Colombia
mjaramillu@eafit.edu.co

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

ABSTRACT

The objective of this report is to analyze and propose a possible solution to the late detection of Roya, one of the most catastrophic plant diseases in history, present in coffee crops in several Latin American countries, including Colombia. In order to do so, an algorithm will be developed that through the study of data collected by a network of wireless sensors is able to analyze and predict which crops have or are likely to have this fungus. The solution to this problem is of paramount importance to the Colombian peasantry because more than half a million families depend on these crops for their livelihoods. For this reason, it is our responsibility as Colombians to contribute to the development of technologies capable of reducing the impact of this infection in the countryside of our country. Like this one, there are similar problems with solutions that can help us to solve our problems, later we will review some of them in order to find the best possible solution.

1. INTRODUCTION

Colombia, a country recognized for its great variety of crops and the quality of its products abroad, has been constantly threatened by the problem of a fungus that has been affecting one of its most desired products internationally, coffee. With more than 563,000 families approximately, the guild of coffee growers makes possible the export of 13.5 million bags of coffee a year, thus achieving to be the main agricultural export product of the country. However, this product has been going through very critical times due to a pest known as Roya, which due to its late diagnosis, makes it very difficult to treat and thus end up with much of the crops. In search of a solution to this problem, the use of a network of wireless sensors has been proposed to maintain these crops with constant monitoring where physical and chemical data related to the appearance of this fungus will be collected.a

2. PROBLEM

The problem we face is based on creating, through the use of data structures, a system capable of relating the data already studied of Caturra coffee plants achieving to establish parameters and possible causes that make the Rust appear in coffee crops, so as to know beforehand if there is the presence of this fungus in the crop studied. To achieve this purpose will represent a great advance for the Colombian agriculture, achieving this way by means of its implementation in the cultures of Caturra coffee to diminish the high quantity of cultures lost by cause of the Rust.

3. RELATED WORK

3.1 ID3 algorithm

ID3 is an algorithm to generate a decision tree created by Ross Quinlan focused on the search for hypotheses or rules based on a set of examples formed by a series of continuous data called attributes in which one will be the attribute to classify. This, also known as objective, is of binary type, that is, it will have values such as positive or negative, yes or no, valid or invalid, etc. The ID3, based on the previously entered examples, tries to obtain the hypotheses by means of which to classify new instances in positive or negative. Figure 1 shows a decision tree generated by the ID3 algorithm.

3.2 C4.5 algorithm

This algorithm, developed by Ross Quinlan, is an extension of the ID3 algorithm mentioned above. C4.5 constructs decision trees from a set of training data in the same way that ID3 does, using the concept of information entropy. At each tree node, the algorithm chooses a data attribute that divides the set of samples into subsets as efficiently as possible. In this way, the attribute with the highest gain of normalized information is chosen as the decision parameter. In order to recursively divide the data into smaller lists. This algorithm has three base cases:

- First case: All samples belong to the same class, for which the algorithm creates a sheet node for the decision tree telling you to choose that class.
- Second case: None of the characteristics provides any information gain. In this case, C4.5 creates a decision node above the tree using the expected value of the class.
- Third case: Instance of the previously unseen class found. For this case, C4.5 will do the same as the previous case.

Quinlan in search of a better algorithm that the ID3 added to C4.5 some improvements like the management of both continuous and discrete attributes, management of formation data with missing attribute values, management of attributes with different costs and the elimination of the branches that do not help in the solution, replacing them with leaf node.

3.3 CART algorithm

As the name implies, CART is a technique with which classification and regression trees can be obtained. When the target variable is discrete, classification is used; when it is continuous, regression is used. The trees created by the CART algorithm are usually very easy to interpret. These trees use historical data with which you build regression trees that allow you to classify and predict new data. In general, this algorithm finds the independent variable that best separates our data into groups, expressing it as a rule to assign its corresponding node. Then, for each of the resulting groups, the same process is repeated recursively until it is not possible to obtain a better separation.

3.4 CHAID algorithm

CHAID, or Chi-squared Automatic Interaction Detection, is a classification method for generating decision trees by chi-square statistics to identify optimal divisions. It was proposed by Gordon V. Kass in 1980 and is currently one of the most used in marketing studies.

4. REFERENCES

[1] CropLife Latin America. Roya del Cafeto. `https://www.croplifela.org/es/plagas/listado-de-plagas/roya-del-cafeto`

[2] Charris, L Henríquez, C, Hernández, S, Jimeno, L, Guillen, O, Moreno S. *Análisis comparativo de algoritmos de árboles de decisión en el procesamiento de datos biológicos.*

[3] Centro de Estudios y Aplicaciones Logísticas. Faculty of Engineering of the National University of Cuyo. *Algoritmo ID3.*

[4] Bosco, J. *Árboles de decisión con R clasificación.*

[5] IBM. *Nodo Chaid.*

[6] Quinlan, J. R. *C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.*