COFFEE LEAF RUST SOLVING: WIRELESS SENSOR NETWORK, DATA STRUCTURES AND ALGORITHMS APPLICATIONS

Pablo Buitrago
Universidad EAFIT
Colombia
pbuitragoj@eafit.edu.co

Miguel Ángel Correa Manrique
Universidad EAFIT
Colombia
macorream@eafit.edu.co

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

ABSTRACT

Humanity presents diverses hazards linked to the agriculture production, sometimes leaded to the distribution or production process, when the last one is a little more difficult to embrace because of the fact that is a natural growth process handed by people or machines.

An example of what was mentioned is the coffee leaf rust (CLR) disease presented in cultivation caturra coffee, a disease that is devastating susceptible coffee plantations, farms have to manage how to solve this problem such as putting hybrids that have strong genetic resistance to rust.

Another way to prevent this is finding a way to detect these diseases by technology prompt resources that could detect them before the mitigation of all plant plantations, increasing the plant survival rate and of course the production amount on farms leaded to resources generated to feed humanity.

1. INTRODUCTION

Nowadays humanity faces difficulties against different ways in how to get resources to feed themselves, particularly with those related with agriculture ambits.

Studies have found that the one majority global warming mitigation comes from meat production and distribution, what leads to think about alternatives such as seeding that could reduce this carbon footprint.

Encountering contrivances that help to do seeding will contribute to the progress of the alternatives spoken such as the ways that could diminish diseases while the seeds grow.

The project has been focused in the CLR disease detection and resolving by the develop of a system artefact that could reduce this disease on caturra coffee, identifying the relevant data needed.

2. PROBLEM

EAFIT University has detected adversities in the growth process of caturra coffee, topic related with the CLR disease, evidencing that having an artificial artefact that could recognize diseases will increase the survival rate on a plantation, meaning it is needed to develop this system and artefact recognition solution to the caturra coffee cultivation disease by qualifying each of them, decision which is derived by the data and information collected from each plant, such as temperature, humidity, illumination, ph, etc... in the plantation.

3. RELATED WORK

3.1 Random Forest

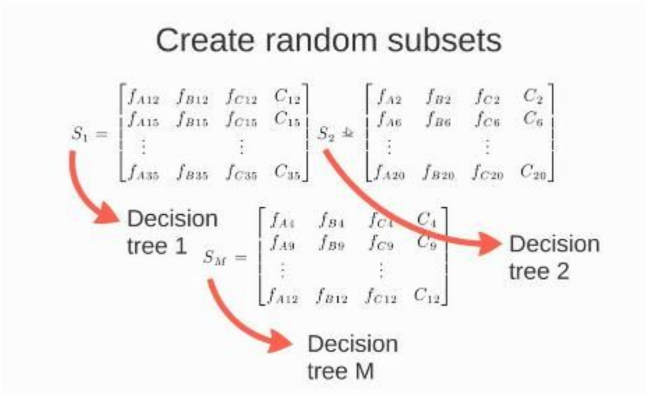
It stands random forest because of the fact that it is used a lot of decision trees.

Random forest algorithm works as a large collection of decorrelated decision trees and it used them to make a classification.

$$S = \begin{bmatrix} f_{A1} & f_{B1} & f_{C1} & C_1 \\ \vdots & & \vdots & \\ f_{AN} & f_{BN} & f_{CN} & C_N \end{bmatrix}$$

The matrix S is a matrix of training samples that were submitted to the algorithm to create a classification model, in S f is defined as feature, each arrow is defined as a sample and C stands for all the other features, meaning that it is already obtained a training class.

The aim is to create a random forest to classify the sample set. In order to achieve this is necessary to create random subsets with random values, where for each subset a decision tree will be created and depending on their evaluation it is going to classified the sample.

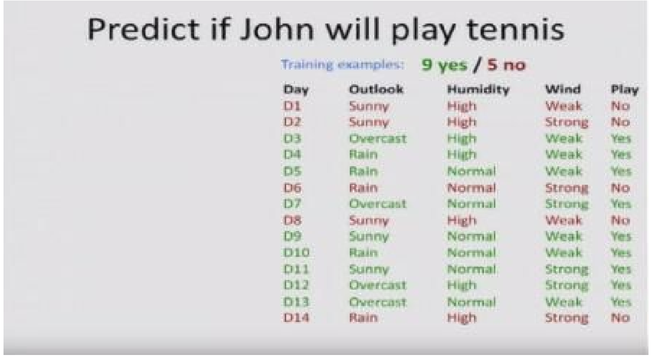


Create random subsets

$$S_1 = \begin{bmatrix} f_{A12} & f_{B12} & f_{C12} & C_{12} \\ f_{A15} & f_{B15} & f_{C15} & C_{15} \\ \vdots & & \vdots & \\ f_{A35} & f_{B35} & f_{C35} & C_{35} \end{bmatrix} \quad S_2 = \begin{bmatrix} f_{A2} & f_{B2} & f_{C2} & C_2 \\ f_{A6} & f_{B6} & f_{C6} & C_6 \\ \vdots & & \vdots & \\ f_{A20} & f_{B20} & f_{C20} & C_{20} \end{bmatrix}$$

$$S_M = \begin{bmatrix} f_{A4} & f_{B4} & f_{C4} & C_4 \\ f_{A9} & f_{B9} & f_{C9} & C_9 \\ \vdots & & \vdots & \\ f_{A12} & f_{B12} & f_{C12} & C_{12} \end{bmatrix}$$

3.2 ID3



Predict if John will play tennis

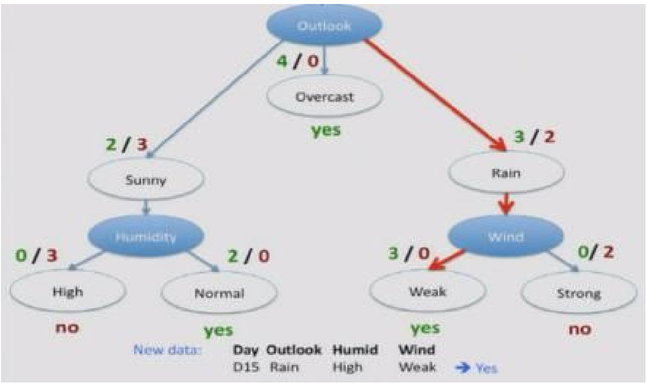
Training examples: 9 yes / 5 no

Day	Outlook	Humidity	Wind	Play
D1	Sunny	High	Weak	No
D2	Sunny	High	Strong	No
D3	Overcast	High	Weak	Yes
D4	Rain	High	Weak	Yes
D5	Rain	Normal	Weak	Yes
D6	Rain	Normal	Strong	No
D7	Overcast	Normal	Strong	Yes
D8	Sunny	High	Weak	No
D9	Sunny	Normal	Weak	Yes
D10	Rain	Normal	Weak	Yes
D11	Sunny	Normal	Strong	Yes
D12	Overcast	High	Strong	Yes
D13	Overcast	Normal	Weak	Yes
D14	Rain	High	Strong	No

It is a recursive algorithm which is going to be divided into "nodes", every point in the algorithm were the data is split because of the fact that one have a lot of training examples with different values on each attribute.

For some attribute A the algorithm will evaluate for the root node if it is already in a pure or not subset, defining them as when while the attribute is given for some classification the attribute is embrace by just one classification or if it is mixed with others respectively.

Being the first case the algorithm stops, it already knows what is the classification for the sample, but if not it will continue with the next child node, evaluating the next attribute given on the sample, repeating the process of evaluation with the pure or not subsets for each child node.

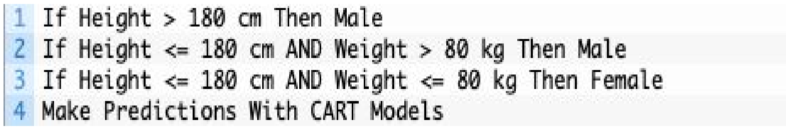


It takes a training set of data and it is going to be sorted into pure subsets based on some attribute values.

3.3 CART

Classification And Regression Trees, or commonly known as CART , is a decision tree algorithm that can be used for classification or regression predictive modeling problems.

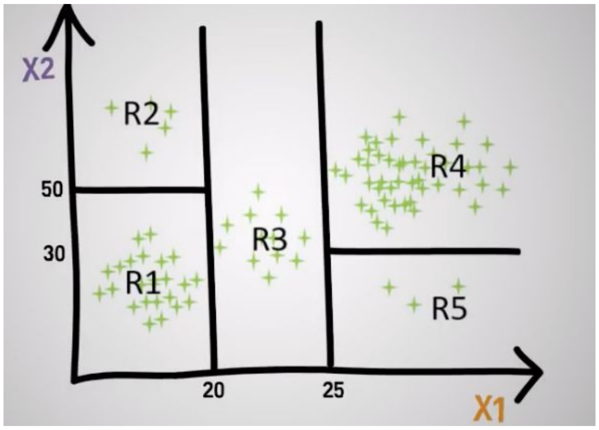
It is a binary tree where each root node represents a single input variable and a split point on that variable (assuming that is numeric); The leaf nodes contain an output variable which is used to make a prediction.



```
If Height > 180 cm Then Male
If Height <= 180 cm AND Weight > 80 kg Then Male
If Height <= 180 cm AND Weight <= 80 kg Then Female
Make Predictions With CART Models
```

(Binary representation of a CART model example)

Given a new input, the tree is traversed by evaluating the specific input started at the root node of the tree.



(image taken from youtube)

A learned binary tree is actually a partitioning of the input space. You can think of each input variable as a dimension on a p-dimensional space. The decision tree split this up into rectangles (quadrants) where all the data will be accommodated once is filtered by the algorithm . It is usually used with greedy splitting (the best split is always selected). The Gini index is used for classification, providing an indication of how "pure" the leaf nodes are.

$$G = sum(pk * (1 – pk))$$

Where G is the Gini index over all classes, pk are the proportion of training instances with class k in the rectangle of interest. A node with perfect class purity will have G=0, and a node with the worst purity will have G=0.5

The most common stopping procedure is to use a minimum count on the number of training instances assigned to each leaf node. If the count is less than some minimum then the split is not accepted and the node is taken as a final leaf node.

3.4 CHAID

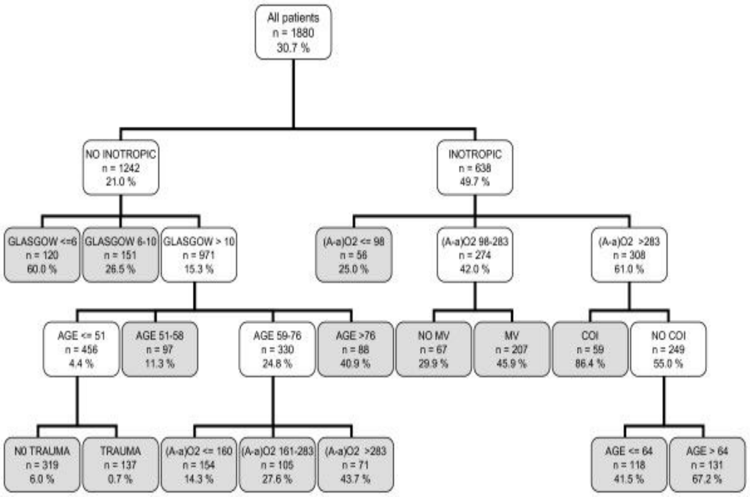
Chi-squared automatic interaction detection or commonly known as CHAID, CHAID will "build" non-binary trees (i.e., trees where more than two branches can attach to a single root or node), based on a relatively simple algorithm that is particularly well suited for the analysis of larger datasets, which for classification problems (when the dependent variable is categorical in nature) relies on the Chi-square test to determine the best next split at each step (https://en.wikipedia.org/wiki/Chi-squared_test) .

The algorithm proceeds as follows:

Preparing predictors. The first step is to create categorical predictors by dividing the distributions, with an approximately equal number of observation into a number of categories.

Mergin categories. Then, the step is to cycle through the predictors to determine for each predictor, the pair of categories that is at least significantly different with respect to the dependent variable. For classification problems with categorical dependent variable it will compute the Pearson Chi-square test. If the test for a given pair of predictor categories is not statistically significant as defined by an alpha to merge value, it will merge the respective predictor categories and repeat this step (find the next pair of categories, which now may include previously merged categories);

Selecting the split variable. The next step is to choose the split variable with the smallest adjusted p-value (the predictor variable that will yield the most significant split). If the smallest adjusted p-value is greater than some alpha to split value, then no splits will be performed and the respectictive node is a terminal one (This process will continue until no further splits can be performed with the given alpha values).


All patients n = 1880 30.7 %
NO INOTROPIC n = 1242 21.0 %
INOTROPIC n = 638 49.7 %
GLASGOW <=6 n = 120 60.0 %
GLASGOW 6-10 n = 151 26.5 %
GLASGOW > 10 n = 971 15.3 %
(A-a)O2 <= 98 n = 56 25.0 %
(A-a)O2 98-283 n = 274 42.0 %
(A-a)O2 >283 n = 308 61.0 %
AGE <= 51 n = 456 4.4 %
AGE 51-58 n = 97 11.3 %
AGE 59-76 n = 330 24.8 %
AGE >76 n = 88 40.9 %
NO MV n = 67 29.9 %
MV n = 207 45.9 %
COI n = 59 86.4 %
NO COI n = 249 55.0 %
N0 TRAUMA n = 319 6.0 %
TRAUMA n = 137 0.7 %
(A-a)O2 <= 160 n = 154 14.3 %
(A-a)O2 161-283 n = 105 27.6 %
(A-a)O2 >283 n = 71 43.7 %
AGE <= 64 n = 118 41.5 %
AGE > 64 n = 131 67.2 %

(image taken from google)

CHAID tree classification example

REFERENCES

[1.Abrams, A. How Eating Less Meat Could Help Protect the Planet From Climate Change. *Time*, 2019. https://time.com/5648082/un-climate-report-less-meat/ .

[2.Brownlee, J. Classification And Regression Trees for Machine Learning. *Machine Learning Mastery*, 2019. https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/.

[3.EAFIT, U. Eafitenses aprovechan la inteligencia artificial para diagnosticar la roya del cafeto. *Eafit.edu.co*, 2019. http://www.eafit.edu.co/noticias/agenciadenoticias/2019/eafitenses-aprovechan-inteligencia-artificial-para-diagnosticar-roya-cafe.

[4.Gupta, B., Rawat, A., Jain, A., Arora, A. and Dhami, N. Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications, 163 (8), 15-18 https://pdfs.semanticscholar.org/fd39/e1fa85e5b3fd2b0d000230f6f8bc9dc694ae.pdf.

[5.Statsoft. Popular Decision Tree: CHAID Analysis, Automatic Interaction Detection. *Statsoft.com*, 2019. http://www.statsoft.com/Textbook/CHAID-Analysis.