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 project is to analyze and propose a possible solution to the late detection of Coffe Leaf Rust, 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 implementation is proposed, 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 economy because more than half a million families depend on these crops for their livelihoods. For this reason, it is our responsibility to contribute to the development of technologies and methods capable of reducing the impact of this infection in the countryside of our country.

Categories and Subject Descriptors
F.2 [**Design and analysis of algorithms**]: Data structures design and analysis

General Terms
Coffee Rust, Roya

Keywords
Pattern matching, Data structures design and analysis

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 presence of a fungus that has been affecting one of its most internationally desired products: coffee. Composed of 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 coffee 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, is very difficult to treat. In search of a solution to this problem, a network of wireless sensors was implemented to maintain a set of coffee crops with constant monitoring, where physical and chemical data related to the appearance of this fungus was be collected. The task at hand is therefore to implement efficient algorithms and data structures that allow classification and understanding of the collected data.

2. PROBLEM
The problem we face is based on creating, through the use of data structures and algorithms, a system capable of finding patterns in already studied data of Caturra coffee plants, as an attempt 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 studied crop. To achieve this purpose will represent a great advance for the Colombian agriculture, through its implementation in the cultures of Caturra coffee to diminish the high quantity of cultures lost by cause of the Rust.

3. RELATED WORK
Our proposed solution is based around the concept of decision trees. We present the following related work as the base material for the construction of our own solution to the problem.

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.

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.

3.3 CART algorithm

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. 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 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. PROPOSED ALGORITHM

The prosed algorithm is based on the same "information gain" principle of C4.5. It can produce both individual trees and random forests.

4.1 Training

1. The data is loaded from the disk, along with information of the nature of each column (called "features"). This includes the feature that the algorithm will try to predict.
2. An untrained tree node is created and the dataset from step 1 (combination of data and feature information) is assigned to it tree.
3. For each feature in the dataset, and for each possible value of it, the information gain is calculated. The feature and value with the best information gain is chosen as the split for this tree node. Additionally, the feature becomes blacklisted for this node, meaning all children nodes will ignore this particular feature when trying to obtain a split. If no information was gained with any of the features, the tree node becomes a leaf.
4. If the tree node did not become a leaf and a split was found, the dataset itself is divided in two parts: The lower part contains all entries where the value of the feature is lower than the split value. The upper part contains the entries where this value is greater or equal.
5. Two untrained tree nodes are created as the left and right children of the original tree node. The lower part of the dataset is assigned to the left, and the upper part is assigned to the right.
6. Step 3 and onwards is repeated for both children until leaves are reached.

4.2 Prediction

1. A dataset entry is extracted from a dataset.
2. If the tree node used is a leaf, the prediction is returned depending on the distribution of the dataset associated to the node. Otherwise, the value of the entry in the feature specified by the tree node is compared to the split value. If the entry value is lower than the one specified by the node, this step is repeated for the left child. Otherwise, this step is repeated for the right child.

4.3 Random forest building

1. A set of untrained trees is created. A random set of features is blacklisted from each tree. Then, a dataset is loaded for each one of them, and then they are individually trained.
2. To predict using a random forest, a prediction is done using each of the trees in the forest. The prediction of each tree is stored to then select the one with the most "votes" (number of trees with the same prediction).

5. DESIGN OF THE DATA STRUCTURE

More than just trying to provide a solution to the main problem itself, we allow one of our goals to be that the implementation of the final algorithm is optimized to allow high volumes of data input. This implies designing a data structure, that the algorithm can use efficiently, such that it takes into consideration the nature of the operations performed.

5.1 Building blocks of the data structure

We present the following *'building blocks'* as the set of basic abstract units that allow the representation of the proposed data structure, as well as an explanation of each one of them:

- `DataValue`: A union type that can hold both discrete (as `int`) and continuous (as `double`) values.
- `DatasetFeature`: [figure 1] A structure that holds the relevant information required to describe a feature in the dataset, along with the name of that feature.

Figure 1: Visual representation of the `DatasetFeature` element.

- `DatasetHeader`: [figure 2] A structure that describes the features of the dataset, as well as the label that the algorithm will seek to optimize.

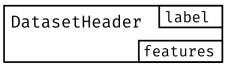

Figure 2: Visual representation of the `DatasetHeader` element.

- `DatasetEntry`: [figure 3] A node in the `Dataset` linked list. It contains an ordered, fixed-size, set of `DataValue` elements. The size of the set is determined by the number of features in the `DatasetHeader` of the associated `Dataset`.
- `Dataset`: [figure 5] This is the main substructure. It is essentially a doubly linked list that holds references to the both the head and tail of a chain composed of `DatasetValue` objects. This structure also has one reference to a `DatasetHeader`. Some relevant properties arise from the design of this block:

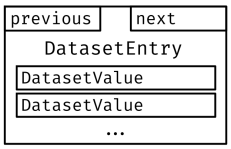

Figure 3: Visual representation of the `DatasetEntry` element.

- Fast [$O(1)$] insertion with a reference. This allows loading data from disk with the maximum possible performance.
- Full iteration capability. No random access is required, which means that the linked list nature of the structure does not affect the overall performance.
- Definition of sub-datasets without increasing the memory usage. As seen in figure 4, two `Dataset` objects can have `head` and `tail` elements in the same chain, thus allowing the definition of sub-datasets without memory redundancy.

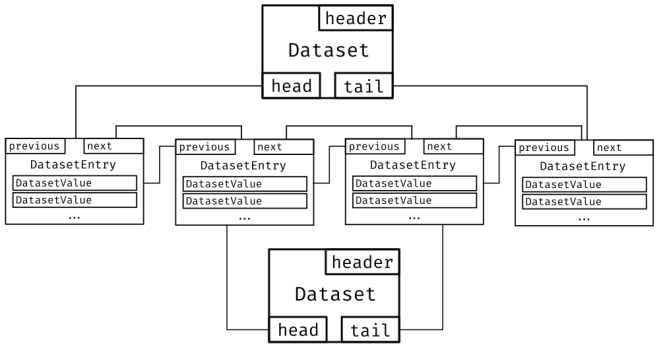

Figure 4: Visual representation of the *sub-dataset* concept.

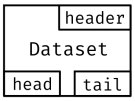

Figure 5: Visual representation of the `Dataset` element.

- `DecisionTree`: A node of a decision tree. It has a reference to a dataset, and two references for the left and right children. It also holds the split value and feature.
- `RandomForest`: A set of decision trees and their associated weights.

5.2 Example data structure usage

Consider a dataset with the features shown in table 1 and the data shown in table 2. Using the proposed data structure, this example dataset would have a similar appearance to that described in figure 6.

name	type	label
fruit	discrete	yes
red	continuous	no
green	continuous	no
blue	continuous	no

Table 1: Features of the example dataset.

fruit	red	green	blue
1 (apple)	246.45	10.32	1.23
1 (apple)	235.23	30.75	3.94
2 (orange)	217.24	184.23	11.23
2 (orange)	250.01	120.88	0.42

Table 2: Data of the example dataset.

5.3 Complexity analysis of the data structure

An analysis of data structure's time complexity provides further information on it's fitness for solving the task. Table 3 presents the worst case time complexity for common tasks that the algorithm performs, as well as an explanation on the parameters for said complexity.

5.4 Data structure benchmarking

We provide some additional statistics on the exection time and the memory usage of the `C` implementation of the data structure:

- Table 4 and figure 7 show the memory (Kb.) usage on dataset tasks.
- Table 5 and figure 8 show the time (Seconds) used during dataset tasks.

6. RESULTS

Using the proposed algorithm and data structure, an accuracy of over 62% was achieved in the identification of Coffe Leaf Rust from the sensor data. There was a mere 2% gain in accuracy from the implementation of random forests. It must be noted that this method of identification of the disease is inherently different to the traditional one, where direct, visul, information is utilized. Therefore, these results have massive meaning, because they show that there exists at least some relationship between the enviroment of the plant and the probability of it getting the disease.

7. CONCLUSION

The use of algorithms for finding patters in biological data is certainly a powerful way to improve the efficiency and quality of products in the modern world. Decision trees are inherently limited, but they provide a cheap and effective method of classification that allows implementation on a big scale. The reduction of complexity in the whole workflow (data collection and classification) has the potential, in this case, to provide significative development on the economy of the country.

For future work, we consider it would be possible to integrate additional data into the algorithm for full automation. For instance, image data of the plants could be transformed into numerical features that this algorithm can utilize to produce

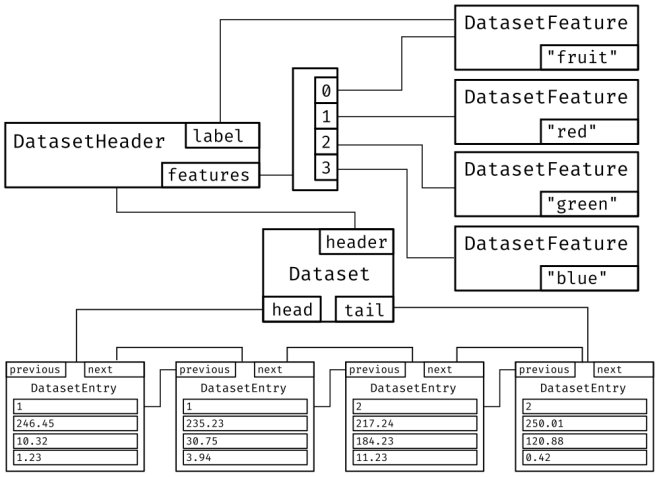

Figure 6: Example dataset appearance.

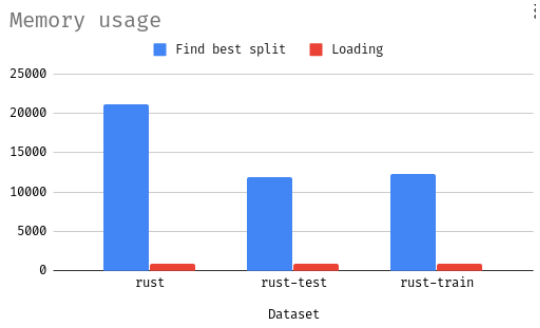

Figure 7: Memory used on dataset tasks.

a better prediction, ultimately reducing even more the need for the human factor.

8. ACKNOWLEDGEMENTS

This research was partially supported by the Colombian Government *Generación E* program.

9. REFERENCES

[1] CropLife Latin America. Roya del Cafeto. [Online]. Available: https://bit.ly/32SYSmZ

[2] L. Charris, C. Henríquez, S. Hernández, L. Jimeno, O. Guillen, S. Moreno, "Análisis comparativo de algoritmos de árboles de decisión en el procesamiento de datos biológicos", *Revista I+D en TIC*, vol. 9, pp. 26-34, 2019.

[3] Bosco, J. (2018, April 23). Árboles de decisión con R - clasificación. [Online] Available: https://bit.ly/2MRvfNs

[4] IBM. Nodo CHAID. [Online] Available: https://ibm.co/32VHLkx

[5] J. R. Quinlan. "C4.5: Programs for Machine Learning". *Machine Learning*, vol. 16 pp. 235-240, September 1994.

[6] N. I. Sujan. (2018, June 29). "What is Entropy and why Information gain matter in Decision Trees?". [Online] Available: https://bit.ly/34dU1Np

Task	complexity	parameters
Loading from disk	$O(n)$	• n: The number of entries to be loaded from disk.
Find best information gain split	$O(mno)$	• m: The number of entries in the dataset. • n: The number of features in the dataset. • o: Variable proportional to the precision used for split evaluation.
Train decision tree	$O(n^3 * mo)$	• m: The number of entries in the dataset. • n: The number of features in the dataset. • o: Variable proportional to the precision used for split evaluation.
Build full random forest	$O(2^n * n^3 * mo)$	• m: The number of entries in the dataset. • n: The number of features in the dataset. • o: Variable proportional to the precision used for split evaluation.

Table 3: Table of complexity for common tasks.

Dataset	Find best split	Loading
rust	21196	964
rust-test	11920	964
rust-train	12300	964

Table 4: Memory used on dataset tasks.

Dataset	Find best split	Loading
rust	3.3113	0.00229582
rust-test	0.2223	0.00109157
rust-train	0.2233	0.00108636

Table 5: Time used on dataset tasks.

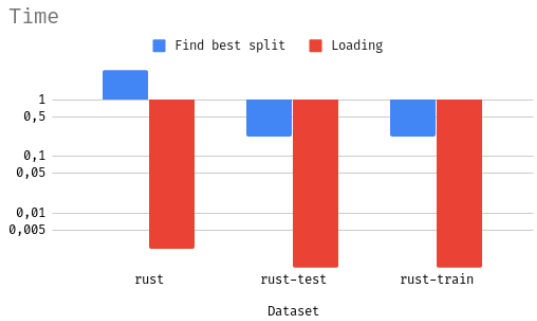

Figure 8: Time used on dataset tasks (logaritmic scale).