LA DETECCIÓN DE LA ROYA EN PLANTAS DE CAFÉ CATURRA

Martín Sánchez Reyes
Universidad Eafit
Colombia
msanchezr@eafit.edu.co

Juan Martín Uribe
Universidad Eafit
Colombia
jmuribef@eafit.edu.co

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

RESUMEN

Según la organización CropLife Latin America [6], la roya del café es el principal problema fitosanitario para la caficultura, y está entre las siete plagas de plantas que ha dejado mayores pérdidas en los últimos cien años. Esta problemática motiva el desarrollo de un proceso analítico para aprovechar un grupo de datos sobre terreno y ambiente para predecir efectivamente el desarrollo del hongo en una determinada planta o área estudiada.

El desarrollo efectivo de este algoritmo es de vital importancia para la sostenibilidad económica de la producción de café, el agro en Colombia y aumentar la competitividad de este producto de exportación.

Palabras clave de la clasificación de la ACM

Opción 1: CCS → Theory of computation→Theory and algorithms for application domains→Database theory→**Data structures and algorithms for data management**

1. INTRODUCCIÓN

En especial, el problema de la roya del café es agravado por la detección tardía de la presencia de la plaga en los cultivos. Cuando ya se diagnostica, es demasiado tarde en la mayoría de los casos para actuar, generando pérdidas que pueden llegar a superar la producción y llevar a la quiebra a muchas familias campesinas. Al llevar a un invernadero y recolectar datos controlados para las plantas, podemos cuantificar la incidencia de las variables medidas para tomar decisiones sobre ellas.

2. PROBLEMA

El problema que nos enfrentamos desde nuestra área, es el desarrollo de un algoritmo que permita, por medio de árboles de decisión, la predicción oportuna de la presencia de la roya bajo las variables de humedad y temperatura del ambiente y del suelo, pH e iluminación. Este problema refuerza la conexión de nuestra área de conocimiento con nuevas aplicaciones y potenciales, al generar beneficio en esta actividad económica del agro en las maneras que ya mencionamos.

3. TRABAJOS RELACIONADOS

3.1 Árboles CHAID

Los árboles CHAID, o de detección automática de interacciones fueron desarrollados en Sudáfrica y posteriormente publicados en 1980 por Gordon Kass. Esta técnica puede ser usada para predecir y detectar interacciones entre variables. Su aplicación se enfoca en el análisis de comportamiento de variables y cómo esta afecta a otras, basándose en esto para predecir demás interacciones que involucran las mismas variables. Una ventaja de la técnica CHAID es que es sumamente visual y fácil de interpretar, y necesita trabajar con grandes cantidades de información para que su análisis sea de mayor confiabilidad. Los árboles CHAID se basan en identificar valores específicos de un valor para determinar si es necesario dividir nodos mediante un nodo de decisión, y a diferencia de otros árboles pueden hacer múltiples divisiones, no solamente binarias. Una ventaja respecto a otros árboles de regresión es que es no paramétrico, lo cual indica que el tamaño del árbol irá cambiando su tamaño dependiendo de la información con la cual esté asociado, y no se está fijando su tamaño desde un inicio lo cual lo hace más flexible en cuanto al conjunto de datos e información con la que se está trabajando.

3.Árboles CART

CART es un concepto más general en cuanto a los árboles de decisión, sumamente importante para modelar algoritmos de predicción. CART, lo cual indica árboles de regresión y clasificación, es un modelo binario de decisiones donde cada nodo representa una única variable de entrada, a diferencia de CHAID, donde se puede hacer múltiples particiones. Estos árboles luego pueden ser representados como gráficas e incluso reglas para la predicción basándose en las variables de entrada. La modelación con datos a menor escala es bastante sencilla, puesto que se basa en decisiones binarias que finalizan en nodos terminales.

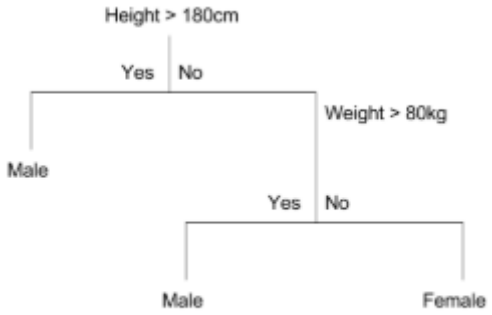


Figura 1[4].

De igual forma a la técnica CHAID, las reglas para hacer una partición entre nodos se basan en el valor de una única variable, y la predicción para un objetivo variable es cada nodo terminal. A partir de estas sencillas características se definen los árboles de regresión y clasificación, donde su complejidad se ve reflejada en la cantidad de nodos y subnodos que el árbol representa.

3.3 Algoritmo ID3

J. Ross Quinlan desarrolló originalmente el ID3 (o el Iterative Dichotomiser 3) en la universidad de Sídney y lo publicó en su libro Machine Learning. Su pseudocódigo es como sigue:

A – Mejor atributo

· Asignar A como atributo de decisión para el nodo.

· Para cada valor de A, crear un descendiente del nodo.

· Clasificar ejemplos de entrenamiento a las hojas.

· If los ejemplos están perfectamente clasificados:

- PARAR

· Else:

- Iterar sobre hojas.

Es un algoritmo que para construir un árbol de decisión basado en atributos no categóricos, formando una distribución de probabilidad, cuya información generada, es llamada entropía.

El mejor atributo es aquel que maximiza la ganancia de información y la reducción de entropía. La función de ganancia está expresada con parámetros S del conjunto de ejemplos dados y A un atributo particular. |S| y |Sv| son el número de elementos y cada valor que puede tomar el atributo.

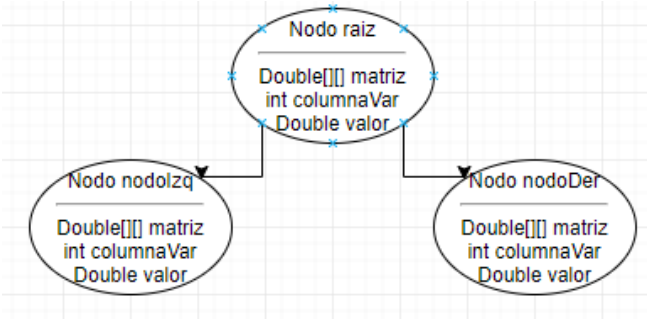
$$Gain(S,A) = Entropy(S) - \sum_{v} \frac{|S_v|}{|S|} Entropy(S_v)$$

3.4 C4.5

Es una extensión del algoritmo original ID3. Es un clasificador estadístico que funciona esencialmente igual que el algoritmo ID3. Analizando, de un conjunto de datos de entrenamiento S= $s_1, s_2, \cdots$, siendo cada s_i un vector p-dimensional, en el que cada entrada es un atributo del dato, incluida la categoría en la que cae. Es un algortitmo recursivo en el que se subdivide el conjunto inicial en cada nodo basado en la diferencia de entroía o ganancia de información.

4. Estructura de datos diseñada

Gráfica 1:



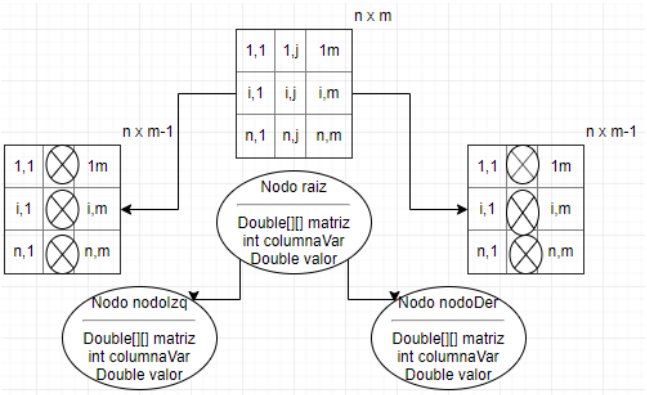
Con la figura 1 representamos la estructura básica del árbol diseñado. Partiendo de un elemento Nodo raiz, definimos como atributos del objeto tanto el valor encontrado como mejor variable, como también la columna en la que se ubica. Los datos entran en una estructura de matriz directamente de la lectura de datos del archivo.

La importancia de esto a nuestro critero radica en definir implícitamente en la estructura la etapa de entrenamiento. Con esto, diferenciarlas dos etapas. En el proceso de prueba, con estos datos establecidos del objeto creado y estructurado, se podrán operar las funciones de clasificación para datos de predicción.

4.1 Operaciones de la estructura de datos

El proceso elegido para la no repetición de variables en la elección de valores de separación es reducir la matriz en una columna, suprimiendo aquella en la que se haya encontrado la variable de separación. Esta operación de la estructura la representamos en la gráfica 2 a continuación

Gráfica 2:



4.2 Criterios de diseñode la estructura de datos

El diseño de la estructura tuvo como criteros, la claridad del proceso algorítmico, pues nuestro desarrollo no se apoyó en librerías o códigos existentes, como también en la facilidad que brinda éste para el paso a la etapa de prueba. La complejidad se vio algo sacrificada, pues el algoritmo debió hacer algunos procesos internos más de una vez. A su vez, la decisión por el árbol CART involucró la versatilidad de la inclusión en código de los criterios que maneja: tanto impureza de Gini, como ganancia de información.

4.3

El análisis de complejidad se divide en lectura, estructuración del árbol y por útlimo, en la etapa de prueba: clasificación.

Complejidad de lectura de datos	O(nxm)
Complejidad estructuración	$O(mxm^2x2^m)$
Complejidad clasificación	O(log m)

6 Conclusiones

Aunque el diseño de la estructura de datos fue completo y se sirvió de los criterios propios al modelo adoptado, hay una gran posibilidad de exploración en términos de eficiencia y capacidad de reducción de complejidad. Aun así, el funcionamiento del código de la estructura resultante es adecuado.

REFERENCIAS

1. Ali, A (2018). Decision Tree (CART) Algorithm in Machine Learning. Recuperado de: https://medium.com/machine-learning-researcher/decision-tree-algorithm-in-machine-learning-248fb7de819e

2. Bennett, K. (n.d.) Global Tree Optimization: A Non-greedy Decision Tree Algorithm.

3. Breiman, L., and J. Friedman, R. Olshen, and C. Stone (1984).*Classification and Regression Trees*, Belmont, CA: Wadsworth International Group.

4. Brownlee, J (2019). Classification and regression tres for machine learning. Recuperado de: https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

5. Colaboradores Wikipedia. (n.d.) C4.5 algorithm. Recuperado de: https://en.m.wikipedia.org/wiki/C4.5_algorithm

6.Instituto Interamericano de Cooperación para la Agricultura (n.d.) Roya del cafeto. Recuperado de: https://www.croplifela.org/es/plagas/listado-de-plagas/roya-del-cafeto

7. Luchman, J (2013). CHAID: Stata module to conduct chi-square automated interaction detection. Recuperado de: https://ideas.repec.org/c/boc/bocode/s457752.html

8. Rao, V. 2013. Introduction to Classification & Regression Trees (CART). Recuperado de: https://www.datasciencecentral.com/profiles/blogs/introduction-to-classification-regression-trees-cart