Detección de la roya por medio de árboles de decisión

Esteban Gonzales Tamayo
Universidad Eafit
Colombia
egonzalezt@eafit.edu.co

David Felipe Garcia Contreras
Universidad Eafit
Colombia
dfgarciac1@eafit.edu.co

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

RESUMEN

El café es unas de las plantas que requiere un cuidado especial para evitar la roya que es una de las enfermedades más nocivas en el café, porque como lo plantea (Consuelo , Oscar , & Roberto) "requiere condiciones climáticas específicas para su producción, su cultivo exige condiciones especiales de suelo, temperatura, precipitación y altitud sobre el nivel del mar", por lo cual cultivar café exige unas condiciones específicas donde es necesario el uso de la tecnología para medir las condiciones de los cultivos y así poder alertar en qué momentos se puede presentar la roya en el café y poder tomar acción temprana a este problema ya que si no se realiza la roya podría generar grandes pérdidas a los caficultores.

Palabras clave

Árboles de decisión; Roya; Algoritmo; Complejidad; Pila; Patrones; Análisis

Palabras clave de la clasificación de la ACM

Mathematics of computing → Information theory→ Coding theory

Theory of computation→ Design and analysis of algorithms→ Mathematical optimization→ Continuous optimization→ Stochastic control and optimization

Theory of computation → Theory and algorithms for application domains → Algorithmic game theory and mechanism design → Network formation

1. INTRODUCCIÓN

La planta de café es una de las plantas más importantes para el ser humano, tanto así que su producción es en masa viéndose a Colombia como una de las tres mayores exportadoras, las cuales se ven afectadas por la plaga de la roya en el café lo cual involucra problemas no solo económicos donde también se ven afectado los cultivadores. Las pérdidas en solo América Latina se ven afectados en un
30% de las cosechas. Retomando lo anterior dicho por lo cual en este trabajo se busca dar una solución con los datos tomados de la roya del café. Se quiere dar una solución óptima en la cual se pueda detectar patrones en la plaga que afecta los cultivos.

2. PROBLEMA

La roya en el café es una situación en la cual muchos cultivadores tienen que pasar por diferentes condiciones. Este problema no se ve como algo nuevo ya que en el año 1868 en la isla de Ceilán se cultivaba el café, pero por culpa de la plaga de la roya cambiaron el café por él te. Si se mueve este problema a Colombia la plaga en los cultivos del café trasciende mucho más ya que si no es tratado se puede ver afectado cerca del 50% del cultivo transformándose este problema en algo más avanzando involucrando factores biológicos. Basándonos en lo anterior se busca a través de un análisis sobre las condiciones de los cultivos de café se pretende detectar patrones o condiciones determinadas que son las causantes de la roya en el café, para poder controlar o prevenir la infestación de este hongo en las plantas de café y así evitar posibles pérdidas materiales en la cultivación del café al igual que la escasez de este cultivo en muchas zonas del país.

3. TRABAJOS RELACIONADOS

3.1 Nodo C5.0

Es un algoritmo usado para generar árboles de decisión, dividiendo información para generar mejores resultados, solamente puede predecir un solo objetivo. El algoritmo como su función principal busca dividir el problema en varias subdivisiones hasta que no hay maneras de dividir luego de realizar esta operación se comienza a eliminar las subdivisiones innecesarias. El algoritmo C5.0 al poder predecir un solo objetivo esta predicción se considera exacta, no solo limitándose a Árboles de decisión por lo que se puede generar además conjuntos de reglas para retener la mayor información posible si se cumple a un registro específico

3.2 Algoritmo C4.5

C4.5 es un algoritmo utilizado para generar árboles de decisión, este se caracteriza porque sus árboles pueden ser utilizado para clasificación estadística. El algoritmo elige cada atributo y considera todas las posibles pruebas para determinar para así poder crear nuevas subdivisiones a partir del atributo base lo cual resulta más eficiente para dividir los datos cuando se analizan los atributos, al igual que este determina con las pruebas cuál dato obtuvo mayor ganancia de información donde este será utilizado como factor o parámetro de decisión. Este algoritmo puede solucionar problemas basados en hechos o condiciones como realizar una actividad ya sea caminar, correr, etc. basándose en los datos que proceso para tomar una decisión de si es conveniente realizarlo o no.

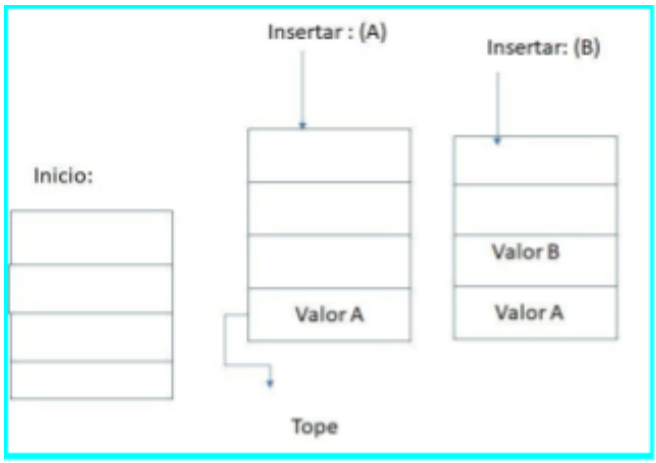
3.3 Algoritmo CART

Es un algoritmo que permite variables de todo tipo, las reglas son de tipo binario permitiendo una estructura de mayor profundidad en los árboles de decisión Para la clasificación se usa la entropía, el índice de Gini, el criterio de Twoing. La función principal de este algoritmo es hallar una estructura de árbol de decisión los más compleja posible, midiendo la complejidad de este por medio de sus nodos terminales, donde se eliminan distintas ramas de manera que combina el riesgo o coste de predicción por lo cual de manera inteligente se usa para eliminar las ramas más débiles, dando así al subárbol más óptimo, permitiendo el alargamiento de los problemas para encontrar por la manera que tome más variables al mismo tiempo , para que su resultado sea el mejor para la utilización dentro del problema.

3.4 Algoritmo ID3

Es un algoritmo que se basa en la búsqueda de una hipótesis o reglas basándose en un conjunto de ejemplos que emplea el algoritmo para determinar sus hipótesis y así poder clasificar todas las instancias basándose en el ejemplo según el tipo de valor que tenga el ejemplo este creará el árbol de decisión. El árbol contendrá ciertos elementos llamados Nodos: Atributos Arcos: valores posibles del nodo principal Hojas: Nodos que clasifican los ejemplos como negativos y positivo.

4. Estructura de datos



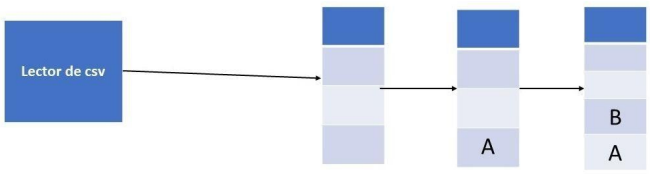
Gráfica 1: Pila.

4.1 Operaciones de la estructura de datos

Como estructura de datos se implementará la estructura de las pilas, para almacenar la información obtenida previamente, la cual será implementada para ser analizada por el árbol, la pila almacena por cada inserción los datos obtenidos en la lectura del archivo csv los cuales son: ph, soil_temperature,soil_moisture,illuminance,env_temperatur e,env_humidity.

Al igual que al momento de analizar la información solamente se necesitará obtener la información

previamente recolectada para luego ser procesada y analizada, sin embargo la información almacenada será almacenada en una copia para trabajar con los datos futuramente.



Gráfica 2: Almacenamiento de datos obtenidos del csv.

4.2 Criterios de diseño de la estructura de datos

Nosotros implementamos esta estructura ya que su complejidad es o(1) para realizar push y pop lo cual es útil para obtener la información procesada del archivo csv para ser procesada en el árbol de decisión para identificar la roya en el café, adicionalmente esta información tendrá una copia de respaldo para conservar la información leída si llega a ser necesitada a futuro, la idea de la pila es enviarle toda la información al árbol para analizarlas y también implementamos esta estructura dinámica precisamente para no tener un espacio fijo de memoria para almacenar la información, si no que se pueda ingresar la información que sea necesaria y también al momento de realizar pop será útil ya que el tamaño o el consumo de la pila disminuye.

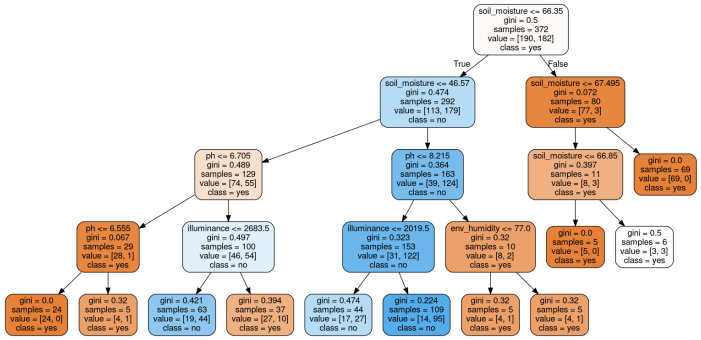
4.3 Análisis de Complejidad

Metodo	Complejidad
Insertar (push)	O(1)
Eliminar(pop)	O(1)
Lectura csv	O(n)

Tabla 1: Tabla de complejidad

5. Árbol de decisión CART

La implementación final del proyecto se implementó la estructura de datos acerca de los árboles de decisión utilizando el el árbol de decisión 'CART' para determinar cuándo un determinado dato posee roya o no, por lo cual este árbol de decisión se basa con el criterio del índice de gini que nos indica que tan mezclados están los datos en un conjunto donde el objetivo del arbol es localizar cual es el dato o variable que mejor parte el árbol y menor índice de gini posee ya que este se encuentra entre 0 y 1 siendo 1 la mayor dispersión de los dato y 0 la separación total de los datos que poseen y no poseen roya.



Gráfica 3: Funcionamiento real del algoritmo Cart

5.1 Operaciones de la estructura de datos

La estructura de datos mencionada anteriormente cuenta con 5 operaciones las cuales son:

Operación	Función
Operacion 1	Lectura de csv
Operacion 2	Creacion del arbol
Operacion 3	Entrenamiento del arbol
Operacion 4	Graficacion del arbol
Operacion 5	Prediccion de los datos

Tabla 2: Definición de las funciones del programa

Cada una de las funciones fueron implementadas para la medición de tiempos en el programa.

- La operación 1 es la encargada de leer los archivos de extensión '.csv'.
- La operación 2 es la encargada de crear el árbol de decisiones 'CART' con el criterio de gini, este posee unos criterios de cuanta profundidad debe tener el árbol y cuantas hojas mínimas debe generar.
- La operación 3 se encargará de entrenar el árbol de decisión y así poder identificar qué variables son las responsables de que haya o no roya.
- La operación 4 se utiliza para generar el árbol gráficamente con los valores con los que fue entrenado.
- La operación 5 se encargará de predecir si un dato en específico posee o no roya y este se basara con los datos previamente cargados para determinar la predicción.

5.2 Criterios de diseño de la estructura de datos

El algoritmo CART se implementó ya que esta se basa con el criterio de Gini lo cual permite saber la impureza de los datos entrados por consiguiente esto permite una facilidad mayor para los datos entrados. Además de lo anterior esto me permite separar los valores y saber donde se tiene que ir las diferentes ramas del árbol para la determinación de la roya junto con esto a la hora de tomar los tiempos se tiene un algoritmo eficiente y capaz de determinar con un solo valor entrado si encuentra Roya o no.

Para esto se requirió de la lectura del CSV por medio de la librería Pandas de Python la cual facilita el análisis de la lectura de los datos por otro lado para la optimización de la graficación del árbol se optó por la librería scikit-learn.

5.3 Análisis de la Complejidad

Operaciones	Complejidad
Lectura del CSV	O(n*m)
Construcción del árbol	O(m*n(log n))
Predicción de los datos	O(n)
Entrenamiento del Árbol	

Archivo Csv n :Filas
m: Columnas

Predicción de los datos: n Cada nodo

Construcción del Árbol: m: Cantidad de las características

n: Cantidad de datos ingresados

Tabla 3: Tabla para reportar la complejidad

5.4 Tiempos de Ejecución

Operacion	Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3
Operacion 1	0,009959221	0,008543015	0,009734392
Operacion 2	0,003386736	0,002781153	0,002662420
Operacion 3	0,004472971	0,004231930	0,005886078
Operacion 4	0,715893269	0,660249472	0,697710276
Operacion 5	0,001644135	0,001250029	0,001026630

Tabla 4: Tiempos en segundos en la ejecución de las operaciones de la estructura de datos con diferentes conjuntos de datos.

5.5 Memoria

Operacion	Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3
Operacion 1	0.7 Mb	0.73 Mb	0.7 Mb
Operacion 2	0.1 Mb	0 Mb	1 Mb
Operacion 3	1 Mb	1.2 Mb	1.4 Mb
Operacion 4	7 Mb	6.6 Mb	7.2 Mb
Operacion 5	0 Mb	0 Mb	0

Tabla 5: **Consumo de memoria de la estructura de datos con diferentes conjuntos de datos**

5.6 Análisis de los resultados

Tabla 8: **Tabla de valores durante la ejecución**

6. CONCLUSIONES

Los árboles como estructura de datos son muy útiles por la simplificación del problema por lo tanto aplicado en el proyecto se puede ver si utilidad aplicada. Basándonos en el algoritmo CART se pudo observar la utilización de técnicas como el índice de gini para poder lograr la estadística de los datos mezclados y su impureza.

Retomando lo anterior en la realización de los datos por medio del csv se pudo ver las variables que más afectan para el entrenamiento del algoritmo para el objetivo de predecir si existe presencia de roya. Al final se cumplió con el objetivo que era lo

que representa el algoritmo CART el cual es la optimización por medio de los nodos.

En cuanto lo que más se dificulta en el algoritmo fue la decisión de cuál era el lenguaje más apto para el reto establecido por lo que se decidió por Python además de esto el cómo influyó el uso de librerías para la facilitación de los cálculos ya que esto nos generó un problema para el entrenamiento del árbol. Por consiguiente para un futuro proyecto se tomará en cuenta una mayor profundidad en cómo es posible lograr el objetivo de una manera simple e óptima por medio de una preparación previa.

REFERENCIAS

CONSUELO MONTES R, OSCAR ARMANDO P, & ROBERTO AMILCAR CADENA, Infestación E Incidencia De Broca, Roya Y Mancha De Hierro en Cultivo De Café Del Departamento Del Cauca, Biotecnología En El Sector Agropecuario y Agroindustrial,2012

Wikipedia. ID3 algorithm, mayo 22, 2019. https://en.wikipedia.org/wiki/ID3_algorithm Amir Ali, 13 Julio 2018
https://medium.com/machine-learning-researcher/decision-tree-algorithm-in-machine-learning-248fb7de819e

Jaime Cárdenas L, Oscar Rodrigo S, & Francisco Orozco M,Royadelcafetero,2018
https://www.croplifela.org/es/plagas/listado-de-plagas/roya-del-cafeto

IBM,NodoC5.0,2019
https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/c50node_general.html

Wikipedia. C4.5, abril 9, 2018. https://es.wikipedia.org/wiki/C4.5

Jorge Martin A, Data Mining con Árboles de Decision
https://web.fdi.ucm.es/posgrado/conferencias/JorgeMartinslides.pdf