ESTRUCTURAS DE DATOS PARA LA DETECCIÓN DE ROYA EN LAS PLANTAS DE CAFÉ

Nayibe Tatiana Vélez David
Universidad EAFIT
Colombia
ntvelezd@eafit.edu.co

Alejandro Fernández Restrepo
Universidad EAFIT
Colombia
afernader@eafit.edu.co

Adelaida Maldonado Esguerra
Universidad EAFIT
Colombia
amaldonade@eafit.edu.co

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

RESUMEN

El objetivo de este informe es analizar y buscar alternativas de solución al problema relacionado con el cultivo de café caturra. Este tiene como objetivo buscar un control a la plaga llamada roya, que atenta con la realización de una buena cosecha y a la pérdida de productos, debido a su tarde diagnóstico. Lo que se busca es encontrar una estructura de datos que permita tomar un monitoreo constante del sembrado para detectar si la cosecha está siendo afectada por la roya.

Para dar solución a nuestro problema, recurrimos a utilizar un árbol de decisión CART, el cual se basa en evaluar cada uno de los datos ingresados y dar a conocer cuál es el más factible a separar los datos que poseen roya. En primer lugar, leemos los datos y procedemos a guardarlos en una matriz, donde se evaluarán cada columna para hallar la variable y el dato con menor impureza, al realizar esto se puede predecir con probabilidades cual planta es factible a tener roya. De esto podemos concluir que, gracias a una estructura de datos, se puede solucionar problemas de la vida cotidiana optando a estos recursos en base a la tecnología.

PALABRAS CLAVES

Estructura de datos, algoritmos, tiempo de ejecución, complejidad, memoria, notación BigO, arboles de decisión.

PALABRAS CLAVES DE LA CLASIDICACIÓN DE LA ACM

Software and its engineering → Software creation and management → Designing Software → Software implementation planning → Software design techniques.

Software and its engineering → Software creation and management → Software verification and validation → Operational analysis

Theory of computation → Design and analysis of algorithms → Data structures design and analysis → Sorting and searching

Theory of computation → Design and analysis of algorithms → Data structures design and analysis → Data compression.

1. INTRODUCCIÓN

En la actualidad, la tecnología se ha convertido en pilar para la humanidad, debido a que con la implementación de esta se busca un avance, logrando que los objetos sean más útiles y propicios para la evolución de la sociedad.

Se puede tomar como ejemplo, la implementación de sensores y estructuras de datos que permitan el control de los cultivos de café caturra, evitando que las plagas, en especial la roya que es la más factible de estar presentes en este tipo de cosechas, sean las causantes el daño y perdida de los productos. De esta manera, lo que se quiere lograr es dar un control y monitoreo constante de lo que está presente en los cultivos a través de la tecnología, que permitirá la evolución y prevención de daño de estos.

2. PROBLEMA

Diseñar un algoritmo con árboles de decisión para la detección de roya en los cultivos de café mediante la obtención de datos del cultivo, tales como: pH, temperatura y humedad del suelo, iluminación, humedad y temperatura del ambiente. El objetivo principal del algoritmo será identificar las variables fisicoquímicas del cultivo para identificar el estado de roya en el mismo.

Este problema deriva otras implicaciones como: la obtención y análisis de los datos, verificar los resultados del algoritmo con el estado actual del cultivo, implementar la solución a casos reales y otros problemas que se tendrán antes y después de la creación del algoritmo. La resolución de este problema permitirá que los grandes cultivos de café tengan un buen desarrollo, el control de plagas y la reducción de la pérdida del producto; facilitando así la exportación de café y seguir consolidando a Colombia como uno de los países con mejor producción de este.

3. TRABAJOS RELACIONADOS

3.1 ALGORITMO ID3

Este algoritmo es basado en la teoría de la computación y tiene como finalidad la construcción de un árbol de decisión, basándose en la clasificación, utilizando el atributo más importante para la selección de los siguientes atributos, basándose en el que posee la información más importante.

[1] Este algoritmo es basado en técnicas de matemáticas y probabilidad y además introduce un concepto: la entropía (medida de incertidumbre); este último es el encargado de elegir el atributo que es más factible a ser seleccionado.

El árbol de decisión está compuesto por nodos: raíz y terminales.

[2] Un problema de ID3 se puede definir por:

- Entradas: Son ejemplos descritos a través de pares(atributo-valor)

Ejemplo	Tipo	Lugar	Estilo	Marco	Autor
E_{17}	grabado	España	moderno	si	A
E_{18}	óleo	Portugal	moderno	no	A
E_{19}	óleo	Francia	moderno	si	B
E_{20}	óleo	España	moderno	no	A
E_{21}	acuarela	España	clásico	no	A
E_{22}	acuarela	Francia	clásico	si	B
E_{23}	acuarela	España	moderno	si	A
E_{24}	acuarela	Portugal	clásico	si	B

- Salidas: Será el árbol construido con la clasificación de los ejemplos a la clase que deben ser parte.

[3] Ross Quinlan define el esquema del algoritmo ID3, el cual adopta los siguientes:

1. Calcular la entropía para todas las clases.
2. Seleccionar el mejor atributo basado en la reducción de la entropía.
3. Iterar hasta que todos los objetos sean clasificados.

Básicamente este algoritmo tiene como finalidad clasificar en orden de importancia la información suficiente para lograr una interpretación de problemas, con el objetivo de minimizar tareas.

3.2 C4.5

Es un algoritmo usado para crear un árbol de decisión, estos árboles pueden ser usados para la clasificación y por eso es referido como un clasificador estadístico.

Este algoritmo fue desarrollado por JR Quinlan y fue una extensión (mejora) del algoritmo ID3. A la hora de crear un árbol de decisión este algoritmo considera todas las posibles pruebas que se pueden dividir del conjunto de datos y selecciona la prueba con mayor ganancia de información.

Este algoritmo se utiliza principalmente en la minería y análisis de datos, aparte este algoritmo

crea y analiza los árboles de decisión de una manera más eficaz que otros algoritmos y reduce la pérdida de datos.

[4] CARACTERÍSTICAS DEL ALGORITMO C4.5
• Permite trabajar con valores continuos para los atributos, separando los posibles resultados en 2 ramas Ai<=N y Ai>N
. • Los árboles son menos frondosos, ya que cada hoja cubre una distribución de clases no una clase en particular.
• Utiliza el método "divide y vencerás" para generar el árbol de decisión inicial a partir de un conjunto de datos de entrenamiento.
• Se basa en la utilización del criterio de proporción de ganancia (gain ratio), definido como I (Xi, C) /H(Xi). De esta manera se consigue evitar que las variables con mayor número de posibles valores salgan beneficiadas en la selección. • Es Recursivo.

3.3 CART

[5] Es un algoritmo de árbol de decisión el cual funciona a través de una o varias entradas de datos y dando como salida un árbol de decisión basado en dichos datos.

Los árboles de decisión son usados mayormente en machine learning (aprendizaje de máquinas), técnicas de modelamiento y clasificación de problemas, también se usan como forma de predicción de datos mediante el análisis y comprobación de una entrada de datos.

Este algoritmo se ha usado principalmente en anuncios y publicidades en las redes sociales a través del análisis de las bases de datos de dichas redes, usando los datos de los usuarios como la edad, el tipo de usuario y hasta pueden llegar a usar el salario promedio de los usuarios.

[6] Otros sectores de aplicación son:
-Industria del seguro
-detección de fraude
-optimización de campañas
-entre otros

CART propone segmentar la base de datos hasta obtener una estructura de árbol lo más compleja posible

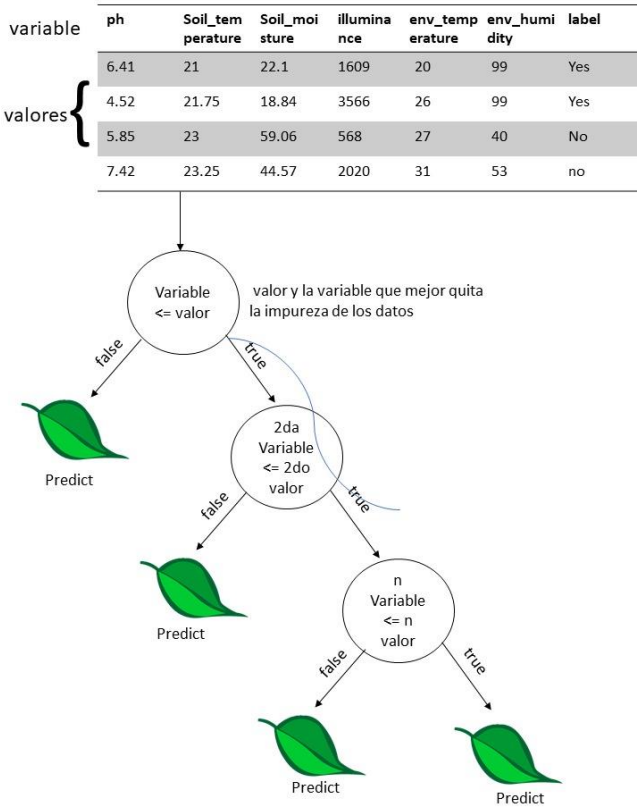
3.4 SLIQ

[7] Sliq es un algoritmo de árboles de decisión usado para la clasificación y análisis de cantidades masivas de datos y también para el minado de datos en bases de datos.
En grandes cantidades de datos la clasificación y el análisis de estos es muy importantes para mantener un orden, de este problema es de donde nace sliq como un algoritmo que solucione estos problemas.

Para la creación de estos árboles el algoritmo principalmente tiene en cuenta dos factores, primero la evaluación y la separación por tipos de los datos y de segundo la creación de particiones de tipos de datos.

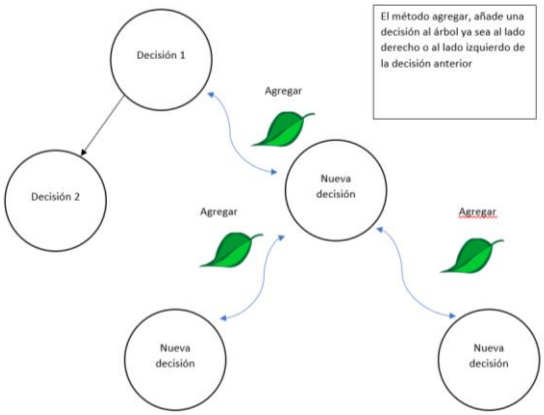
Sliq es un algoritmo que es capaz de manejar tanto como números y diferentes tipos de variables para crear los árboles de decisión sobre estos.

4.TITULO DE LA PRIMERA ESTRUCTURA DE DATOS DISEÑADA



Gráfica 1: Árbol de decisión basado en CART, para dar a conocer la presencia de roya en un cultivo de café

4.1 OPERACIONES DE LA ESTRUCTURA DE DATOS



Gráfica 2: Representación de la operación insertar.

[8] 4.2 CRITERIOS DE DISEÑO DE LA ESTRUCTURA DE DATOS

Hemos elegido, para realizar el algoritmo del árbol de decisión, el modelo CART que se entiende como Classification And Regression Trees. Este modelo puede usarse tanto como para la clasificación o regresión predictiva de modelos. La representación del modelo CART es binaria donde cada raíz representa un input variable y a la vez un punto de división en esa variable. Los outputs nodes del árbol son sus hojas que son usadas para hacer la predicción. La selección de qué variable de entrada usar y el punto de corte se elige usando un algoritmo para minimizar una función. La construcción del árbol finaliza utilizando un criterio de detención predefinido, como un número mínimo de instancias de entrenamiento asignadas a cada nodo de hoja del árbol. Debe haber un orden especial para el orden de las variables pues de lo contrario el árbol no sería efectivo y las respuestas serian diferentes en cada ocasión. Para esto, se utiliza un procedimiento numérico donde todos los valores están alineados y se prueban diferentes puntos de división utilizando una función de costo en este caso gini. Se selecciona la división con el mejor costo (costo más bajo). Así pues, este modelo encaja perfectamente y en lo que queremos realizar en la predicción de la roya donde para saber el orden de las raíces, previamente hallaremos con una función la impureza de cada variable y su respectivo valor el cual también se hallara usando la fórmula de impureza. Después de tener el valor de la impureza de cada variable, se ordenarán de menor a mayor y ese será el orden de las raíces. Al finalizar todo esto cuando se ingrese una nueva planta por el árbol de decisión, se llegará a la predicción de la roya.

4.3 ANÁLISIS DE LA COMPLEJIDAD

MÉTODO	COMPLEJIDAD
Leer	O(n)
Guardar	O(n)

Gráfica 3: Complejidad de los algoritmos de los diferentes métodos.

4.4 TIEMPOS DE EJECUCIÓN

	Archivo
Lectura	10ms

Gráfica 4: Tiempo que se tardan los algoritmos en ejecutarse.

4.5 MEMORIA

	Archivo
Lectura	1 MB

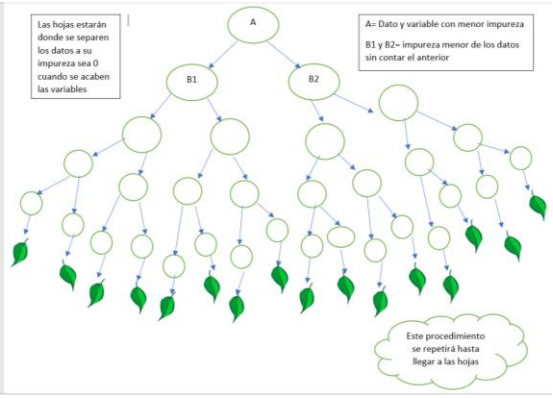
Gráfica 5: Espacio de memoria

4.6 ANÁLISIS DE LOS RESULTADOS

Estructura de autocompletado	Matriz
Recorrido	10 ms
Tamaño	1 MB

Gráfica 6: Unión de los resultados para un análisis previo.

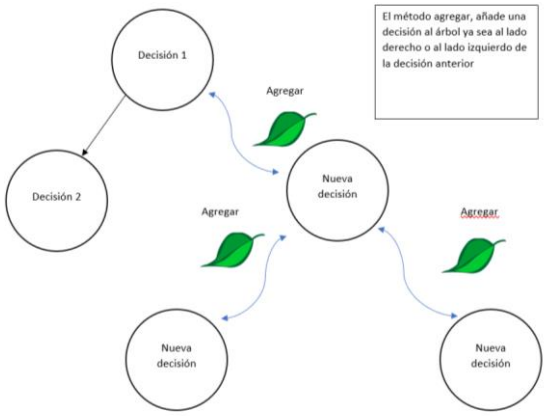
5. TÍTULO DE LA SOLUCIÓN FINAL DISEÑADA



Gráfica 7: Estructura de datos seleccionada: CART

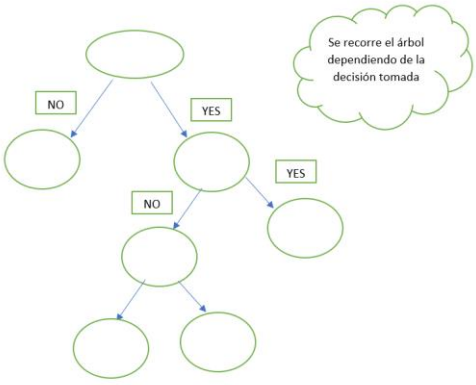
5.1 Operaciones de la estructura de datos

AGREGAR



Gráfica 8: Imagen de una operación agregar en un árbol CART

RECORRIDO



Gráfica 8: Imagen de una operación recorrer en un árbol CART

5.2 Criterios de diseño de la estructura de datos

Inicialmente el manejo de datos en nuestro proyecto lo realizamos a través de una matriz, la cual permite tener un control en el manejo de los mismos, es por esto que al realizar la elección de la estructura de datos nos basamos en que nos permitiera separar los datos de forma óptima y más precisa. Es por esto que elegimos la estructura de datos CART, ya que es una de las que mejor están planteadas y nos permite realizar nuestro objetivo de manera eficiente; otra de las razones por la cual elegimos esta estructura es por que tiene el método de la impureza de Gini, lo cual nos ayuda a la detección de roya en las plantas.

5.3 Análisis de la Complejidad

OPERACIÓN	TIEMPO
Lectura de Datos	O(n)
Evaluar	O(m x o x p)
Impureza	O(q x r)
Creación árbol	O(k)

Tabla 1: Tabla para reportar la complejidad

Variables:

n: Tamaño de lineas del archivo

m:Tamaño matriz

o: Ancho matriz

p: Nivel profundo del árbol

q: Ancho matriz

r: Largo matriz

k: Nivel de profuncidad árbol(datos con impureza)

5.4 Tiempos de Ejecución

OPERACIÓN	TIEMPO
Lectura de Datos	2ms
Evaluar	1ms
Impureza	35ms
Creación árbol	1s

Tabla 2: Tiempos de ejecución de las operaciones de la estructura de datos.

5.5 Memoria

	Conjunto de Datos
Consumo de Memoria	40MB

Tabla 3: Consumo de memoria de la estructura de datos.

5.6 Análisis de los resultados

	Tiempo	Memoria
DataSet train	51ms	38MB
DataSet Test	43ms	30MB

Tabla 4: Tabla de valores durante la ejecución

6. CONCLUSIONES

En base a lo estudiado y procesado en nuestro Proyecto, podemos concluir que a partir de esto podemos estar en capacidad de aplicar las estructuras de datos a la solución de problemas que están en nuestro diario vivir; basicamente nos basamos en árboles que ayudan a tomar la decision frente a una problemática planteada como lo es la roya en las plantas de café a través de los estudios de diferentes variables.

Haciendo énfasis en lo estudiado, pudimos detectar lo que afecta a la planta gracias al uso de la probabilidad, que nos permitió hacer un tanteo de las plantas que son factibles tener roya, logrando con ello, obtener porcentajes que dan como resultado final la información de la presencia de roya o no en un cultivo de café.

A través del desarrollo de nuestro proyecto hemos evolucionado notoriamente, más que todo en la creación de la estructura de datos, logrando importar un DataSet a un árbol de decisión, lo cual al entregar el primer reporte aun no lo teníamos claramente establecido y desarrollado, por lo medio se puede ver un notable avance en el mismo. Pese a que es un tema que es de nuestro dominio, los más grandes problemas se nos presentaron con la creación del árbol para sí proceder con la finalización del programa.

6.1 Trabajos futuros

En nuestro trabajo nos interesa mejorar la eficiencia en la creación automática del árbol, además, el interés por conectar la programación con los sensores para así tener un control total del cultivo, basado en el conocimiento de lo que puede atacar al mismo, así como lo puede hacer la roya. Entrando a detalles de nuestro proyecto, nos gustaría reducir el tiempo de ejecución y el uso de memoria que el mismo proporciona.

AGRADECIMIENTOS

En el proceso de la elaboración de nuestro proyecto contamos con ayuda que permitió el desarrollo del mismo de la mejor manera; es por esto que extendemos un agradecimiento a nuestros monitores, los cuales estuvieron atentos a las dudas que en el transcurso surgieron y muy especialmente damos las gracias a nuestros compañeros, que, al compartir nuestras ideas, ayudaron a establecer de alguna manera una dirección a nuestro proyecto. De antemano, agradecemos a todas las personas que hicieron posible la elaboración de este proyecto.

REFERENCIAS

[1] [3] Wikipedia. Árbol de decisión (modelo de clasificación ID3). Retrieved August 11 2019. From http://bit.ly/2N0pEoF

[2] Fernando Sancho Caparrini. Aprendizaje Inductivo: Árboles de Decisión. 26 of December de 2018. Retrieved August 11 2019. From http://bit.ly/2MZnDJ9

[4] Espino, Tijerina, Cedano, de la Fuente , Pérez, Chiñas. ~ ALGORITMO C4.5 ~. November 2005. Retrieved August 11 2019. From http://bit.ly/2yTgnq2

[5] Amir Ali. Decision Tree (CART) Algorithm in Machine Learning. July 2018. Retrieved August 11 2019. From http://bit.ly/2KIJdyG

[6] Jorge Martín Arevalillo. Data Mining con Árboles de Decisión. Retrieved August 11 2019. From http://bit.ly/2Z1BnWa

[7] Manish Mehta, Rakesh Agrawal and Jorma Rissanen. SLIQ: A Fast Scalable Classifier for Data Mining. Retrieved August 11 2019. From http://bit.ly/2H1rmSD

[8] Jason Brownlee: Classification And Regression Trees for Machine Learning. Retrieved April 8 2016.From http://bit.ly/33dgucX