Detecciòn de la roya en el cafè caturra por medio de àrboles de decisiòn

Manuela Herrera
Universidad EAFIT
Colombia
mherreral@eafit.edu.co

Samuel Palacios
Universidad EAFIT
Colombia
sdpalaciob@eafit.edu.co

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

NOTA DEL DOCENTE: Para ampliar información sobre los requerimientos aquí descritos, consulten la "*Guía para la realización del Proyecto Final de Estructura de Datos 1*" que se entrega. Al final: 1. Borrar este texto escrito en rojo, 2. Adecuar los espacios de los textos, 3. Cambiar el color de los textos a negro. Consideren además que:

Textos en negro = lo que deben hacer en la entrega 1

Textos en Azul = lo que deben hacer en la entrega 2

Textos en violeta = lo que deben hacer en la entrega 3

RESUMEN

El problema a tratar en este proyecto es la tardía detección de la roya en los lotes de café, puntualmente el de tipo caturra, que es de mayor exportación. Si la roya se detecta a tiempo, se pueden prevenir significativas pérdidas de bultos de este, lo que en otras palabras significa impulsar la industria cafetera y elevar los ingresos económicos del país y de las familias caficultoras. Sabiendo que Colombia es un gran exportador, no sólo de café, sino también de flores, caña de azúcar y demás elementos derivados de la agricultura, si se sabe cómo interpretar las variables de los cultivos como el pH, la humedad, entre otros, entonces se podrá aumentar el índice de producción, lo que significa un beneficio tangible en la agricultura y la economía.

/*Para escribirlo pueden dar respuesta a estas preguntas: ¿Cuál es el problema?, ¿Por qué es importante el problema?, ¿Qué problemas relacionados hay?*/, ¿Cuál es la solución?, ¿cuáles los resultados? y, ¿Cuáles las conclusiones? Utilizar máximo 200 palabras.

PALABRAS CLAVE

Detección de la roya, algoritmos en cultivos, árboles de decisión, café caturra, filtración y análisis de datos, redes neuronales, entrenamiento de datos, modelos algorírmicos, estudios de la roya

Palabras clave de la clasificación de la ACM

Information systems > Database design and models > Design and analysis of algorithms > Data structures

1. INTRODUCCIÓN

Colombia es un país que basa su economía en el sector primario; muchas de las familias colombianas dependen de cultivos o del ganado, por lo que, si una plaga como la roya se vuelve corpórea, afecta en gran medida la industria, generando perdidas en su mayoría, significativas.

Hasta el momento la detección de este tipo de plagas está muy mal optimizada, generando así un déficit en el desarrollo de la agricultura, lo que desemboca en no llegar a ser un país competitivo con referencia a la exportación de productos cultivados en el territorio. Por otra parte, los métodos utilizados para el control de dichos factores son poco amigables, el uso de químicos, venenos y demás elementos para el control de plagas hacen que la calidad de los cultivos disminuya, afectando no solo vendedor sino también al consumidor.

Haciendo uso de los árboles de decisión se busca prevenir la pérdida de un sustancial porcentaje de los cultivos, además de reemplazar métodos obsoletos para el control de plagas generando de este modo una alternativa que mejora en todo aspecto el factor comercial e industrial de los agricultores colombianos.

/*Es la justificación de las condiciones en el mundo real que llevan al problema. En otras palabras, es hablar sobre qué va a tratar el documento e incluir la historia de este problema.*/

2. PROBLEMA

En síntesis, la detección tardía de la roya en los cafetales, genera pérdidas inimaginables para la industria colombiana, la disminución en la producción y exportación sumado a las prácticas poco amigables con el cultivo y el consumidor como el uso de químicos y plaguicidas son factores que de ser tratados de forma oportuna pueden llegar a verse reflejados en un mejor posicionamiento del país en cuanto a exportación y aprovechamiento de recursos naturales, además de propiciar el uso de la tecnología como fundamento para el impulso de las familias agricultoras

3. TRABAJOS RELACIONADOS

3.1 Algoritmo ID3

Este algoritmo se encarga de crear el árbol de decisión por medio de un nodo principal y unos terminales, y así, poder generar el árbol mínimo para un conjunto definido. De esta forma, por medio de preguntas se establecen para posteriormente crear un modelo "predictivo" para poder asociar los elementos del conjunto a una clase en particular.

3.2 Algoritmo C5

El algoritmo se encarga de dar un máximo de información sobre un conjunto de reglas sobre las cuales se ejecuta el árbol de decisión, donde sencillamente se muestran las divisiones que hay en el código, de forma tal que por medio del todo que comprenden los nodos y los datos se logre predecir el comportamiento del algoritmo.

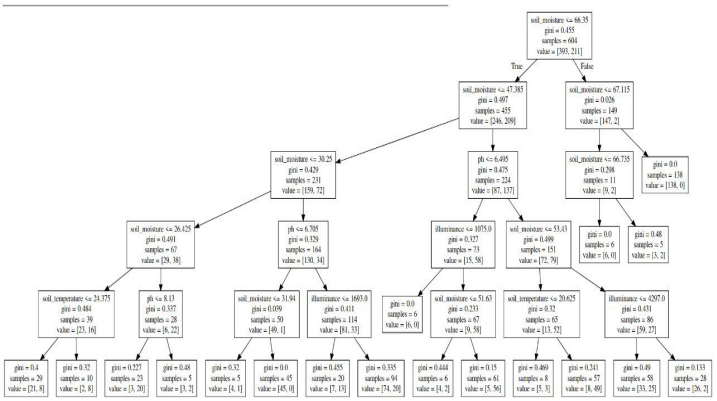
3.3 Algoritmo CART

Conocido como classification and regression tree es un tipo de árbol estrictamente binario que funciona mediante las particiones recursivas. para su uso requiere un data set clasificado y un conjunto de entrenamiento. Sirve para la selección de características de datos tanto numéricos como categóricos utilizando datos históricos para predecir y/o clasificar nuevos datos, para definir que particiones hacer o que reglas utilizar, implementa un criterio de separación Phi "Φ" en base a promedio y sumatorias y también con el índice de Gini en base a las clases probando así todas las reglas de particionado y escogiendo las más optimas en cada caso.

3.4 Algoritmo C4.5

También llamado J48 es una ampliación de ID3. Es de tipo eneario lo que quiere decir que funciona haciendo particiones de forma recursiva, pero sin restricción, a diferencia del binario. Su criterio de división está fundamentado a partir de la ganancia de información, es decir, utilizando la entropía de Shannon y probando en un principio el data set sin reglas estableciendo una entropía inicial posteriormente prueba la entropía de todas las reglas de división y así obtiene el resultado óptimo.

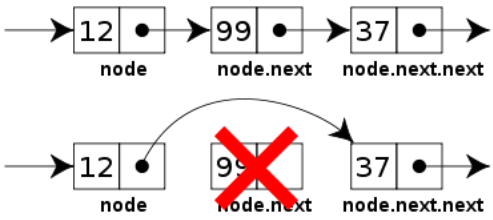
4. Algoritmo CART



Gráfica 1: Árbol de clasificación CART, cada planta tiene las respectivas variables de humedad del ambiente, pH, temperatura del suelo, temperatura del ambiente

4.1 Operaciones de la estructura de datos

Diseñen las operaciones de la estructura de datos para solucionar el problema eficientemente. Incluyan una imagen explicando cada operación



Gráfica 2: Imagen de una operación de borrado de una lista encadenada

4.2 Criterios de diseño de la estructura de datos

Ahora nos plantearemos el por qué de la elección del algoritmo CART; durante la búsqueda de algo que se acomodara a nuestras necesidades vimos diferentes tipos de árboles, pero en particular nos llamó la atención CART porque no teníamos que suponer cosas sobre los datos para trabajar sobre ellos, funciona bien con grandes cantidades de datos, además es fácil entender la resolución del árbol sobre si una planta tiene roya o no, y como plus a lo anterior, resulta más sencilla o intuitiva su implementación.

4.3 Análisis de Complejidad

Operaciones	Complejidad
Construcción del árbol por nodo	O(m(log n)) M: cantidad de características N: cantidad de la muestra
Construcción total	O(m*n(log n))
Lectura y almacenamiento de datos	O(n)
Precisión del subconjunto	O(n)
Split	O(n/2)
validación	O(1)

Tabla 1: Tabla para reportar la complejidad

4.4 Tiempos de Ejecución

	Conjunto de datos 1	Conjunto de datos 2
creación	3.5s	6s

Tabla 2: Tiempos de ejecución de las operaciones de la estructura de datos con diferentes conjuntos de datos

4.5 Memoria

	Conjunto de datos 1	Conjunto de datos 2
Consumo de memoria	15 MG	20 MG

Tabla 3: Consumo de memoria de la estructura de datos con diferentes conjuntos de datos

4.6 Análisis de los resultados

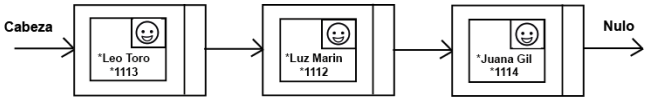
Así, viendo cómo funciona el algoritmo, notamos que consume más memoria y se demora más en ejecutarse el conjunto de datos 2, debido a que tiene mayor número de registros; aún así podemos evidenciar que, con muchos valores, sigue siendo relativamente rápido

	Conjunto de datos 1	Conjunto de datos 2
Tiempo de creación	3.5s	6s
Consumo de memoria	15 MG	20 MG

Table 4: Análisis de los resultados obtenidos con la implementación de la estructura de datos

5. TÍTULO DE LA SOLUCIÓN FINAL DISEÑADA

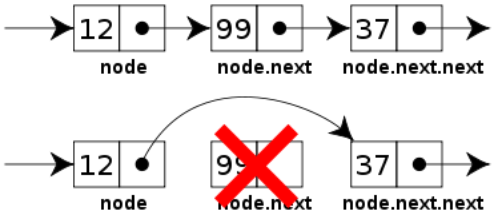
Implementen una estructura de datos para solucionar finalmente el problema y grafíquenla. Además, pruébenla con los datos que están en la carpeta de Conjunto de Datos del .ZIP



Gráfica 3: Lista simplemente encadenada de personas. Una persona es una clase que contiene nombre, cédula y foto

5.1 Operaciones de la estructura de datos

Diseñen las operaciones de la estructura de datos para solucionar finalmente el problema. Incluyan una imagen explicando cada operación



Gráfica 4: Imagen de una operación de borrado de una lista encadenada

5.2 Criterios de diseño de la estructura de datos

Expliquen con criterios objetivos, por qué diseñaron así la estructura de datos. Criterios objetivos son, por ejemplo, la eficiencia en tiempo y memoria. Criterios no objetivos y que rebajan la nota son: "me enfermé", "fue la primera que encontré", "la hice el último día", etc. Recuerden: este es el numeral que más vale en la evaluación con 40%

5.3 Análisis de la Complejidad

Calculen la complejidad de las operaciones de la nueva estructura de datos para el peor de los casos. Vean un ejemplo para reportarla:

Método	Complejidad
Búsqueda Fonètica	O(1)
Imprimir búsqueda fonètica	O(m)
Insertar palabra busqueda fonètica	O(1)
Búsqueda autocompletado	O(s + t)
Insertar palabra en TrieHash	O(s)
Añadir búsqueda	O(s)

Tabla 5: Tabla para reportar la complejidad

5.4 Tiempos de Ejecución

Calculen, (I) el tiempo de ejecución y (II) la memoria usada para las operaciones de la nueva estructura de datos, para el Conjunto de Datos que está en el ZIP. Explicar el tiempo para varios ejemplos

Tomen 100 veces el tiempo de ejecución y memoria de ejecución, para cada conjunto de datos y para cada operación de la estructura de datos

	Conjunto de Datos 1	*Conjunto de Datos 2*	*...Conjunto de Datos n*
Creación	10 sg	20 sg	5 sg
Operación 1	12 sg	10 sg	35 sg
Operación 2	15 sg	21 sg	35 sg
Operación n	12 sg	24 sg	35 sg

Tabla 6: Tiempos de ejecución de las operaciones de la estructura de datos con diferentes conjuntos de datos

5.5 Memoria

Mencionar la memoria que consume el programa para los conjuntos de datos

	Conjunto de Datos 1	*Conjunto de Datos 2*	*...Conjunto de Datos n*
Consumo de memoria	10 MB	20 MB	5 MB

Tabla 7: Consumo de memoria de la estructura de datos con diferentes conjuntos de datos

5.6 Análisis de los resultados

Expliquen los resultados obtenidos. Hagan una gráfica con los datos obtenidos, como por ejemplo:

Tabla de valores durante la ejecución

Estructuras de autocompletado	LinkedList	Arrays	HashMap
Espacio en el Heap	60MB	175MB	384MB
Tiempo creación	1.16 - 1.34 s	0.82 - 1.1 s	2.23 - 2.6 s
Tiempo búsqueda ("a")	0.31 - 0.39 s	0.37 - 0.7 s	0.22 - 0.28 s
Tiempo búsqueda ("zyzzyvas")	0.088 ms	0.038 ms	0.06 ms
Búsqueda ("aerobacteriologically")	0.077 ms	0.041 ms	0.058 ms
Tiempo búsqueda todas las palabras	6.1 - 8.02 s	4.07 - 5.19 s	4.79 - 5.8 s

Tabla 8: Tabla de valores durante la ejecución

6. CONCLUSIONES

Para escribirlas, procedan de la siguiente forma: 1. En un párrafo escriban un resumen de lo más importante que hablaron en el reporte. 2. En otro expliquen los resultados más importantes, por ejemplo, los que se obtuvieron con la solución final. 3.Luego, comparen la primera solución que hicieron con los trabajos relacionados y la solución final. 4. Por último, expliquen los trabajos futuros para una posible continuación de este Proyecto. Aquí también pueden mencionar los problemas que tuvieron durante el desarrollo del proyecto

6.1 Trabajos futuros

Respondan ¿Qué les gustaría mejorar en el futuro? ¿Qué les gustaría mejorar estructura de datos o a la implementación?

AGRADECIMIENTOS

Identifiquen el tipo de agradecimiento que van a escribir: para una persona o para una institución.Tengan en cuenta que: 1. El nombre del docente no va porque él es autor. 2. Tampoco sitios de internet ni autores de artículo leídos con quienes no se han contactado. 3. Los nombres que sí van

son quienes ayudaron, compañeros del curso o docentes de otros cursos.

Aquí un ejemplo: Esta investigación fue soportada parcialmente por [Nombre de la fundación que paga su beca].

Nosotros agradecemos por su ayuda con [una técnica particular o metodología] a [Nombre, Apellido, cargo, lugar de trabajo] por sus comentarios que ayudaron a mejorar esta investigación.

BORRAR LOS CORCHETES ([]).

REFERENCIAS

Referenciar las fuentes usando el formato para referencias de la ACM. Léase en http://bit.ly/2pZnE5g Vean un ejemplo:

1.Adobe Acrobat Reader 7, Asegúrense de justificar el texto. http://www.adobe.com/products/acrobat/.

2. Fischer, G. and Nakakoji, K. Amplifying designers' creativity with domainoriented design environments. in Dartnall, T. ed. Artificial Intelligence and Creativity: An Interdisciplinary Approach, Kluwer Academic Publishers, Dordrecht, 1994, 343-364.

López, B. ALGORITMO ID3. Recuperado de http://www.itnuevolaredo.edu.mx/takeyas/

IBM knowledge center. Nodo C5.0. Recuperado de https://www.ibm.com/support/knowledgecenter/es/SS3RA7_sub/modeler_mainhelp_client_ddita/clementine/c50node_general.html