ALGORITMOS BASADOS EN ÁRBOLES DE DECISIÓN Y EN DATOS PRUEBA SABER 11

Miguel Ángel Martínez Flórez Universidad Eafit Colombia mamartinef@gmail.com	Pablo Maya Villegas Universidad Eafit Colombia pmayav@eafit.edu.co	Miguel Correa Universidad Eafit Colombia macorream@eafit.edu.co	Mauricio Toro Universidad Eafit Colombia mtorobe@eafit.edu.co

RESUMEN

El objetivo de este informe es analizar y solucionar el problema de diseño de un algoritmo cuya principal función es gestionar un algoritmo que, basado en árboles de decisión, pueda predecir si un lote de café caturra está infectado con la roya o no. En Colombia, el café no solo es una de las ramas de exportación agrícolas que más sustento genera en nuestro país, ayudando a más de 563.000 familias. Es así como el tema de las plagas de la roya toma importancia como principal problema fitosanitario que afecta al café, este problema es agravado porque se hace un diagnostico tarde, causando que su control como plaga sea difícil de manejar, inevitablemente, la roya genera grandes pérdidas al negocio del café cada año.

Entre las plagas que amenazan constantemente los lotes de café no solo se encuentra la roya, si no también plagas como: la broca del fruto; el minador de la hoja; piojo harinoso del follaje y de la raíz; barrenador del tallo y araña roja. Corchosis de la Raíz del Cafeto. Estas enfermedades que le causan daño al café son un gran problema, ya que, como mencionamos anteriormente el café es uno de los sustentos más importantes de Colombia.

Palabras clave

Árboles de decisión, cultivo, plagas, predicción de los resultados de un lote de café.

1. INTRODUCCIÓN

Los tres principales productores de café en el mundo son Brasil (con 43.2 M de sacos al año), Vietnam (con 27.7 M de sacos al año) y Colombia (con 13.5 M de sacos al año). En Colombia, el café es la principal exportación agrícola, reconociendo su gran importancia cuando hablamos de reconocimiento y sustento económico. El café colombiano es muy demandado hoy en día por su calidad de sabor, pero para llegar a tener un gran producto, primero se tiene que tener una buena preparación de este, el gran problema que tienen los agrícolas a la hora de preparar el café son las diferentes plagas que se

manifiestan de manera inoportuna arruinando grandes cantidades de cultivos. Por ello es importante encontrar una manera de solventar este problema que a lo largo de los años ha sido uno de los principales problemas para los agricultores colombianos.

1.1. Problema

A través del análisis de algoritmos y árboles de decisión, llegar a implementar un algoritmo que resalte las predicciones en los lotes de café, con el objetivo de predecir qué lote estará infectado con la roya y cual no. A causa de estas predicciones los agricultores podrán tener un mejor manejo a la hora de la producción del café.

1.2 Solución

En este trabajo, nos centramos en los árboles de decisión porque proporcionan una gran aplicabilidad. Evitamos los métodos de caja negra como las redes neuronales, las máquinas de soporte vectorial y los bosques aleatorios porque carecen de aplicabilidad.

Para solucionar el problema utilizaremos un árbol CART, este va a separar los datos según sus

características más importantes y ver cómo estas afectan en la predicción de un lote de café con el fin de responder si un lote de café caturra está infectado con la roya o no.

1.3 Estructura del artículo

En lo que sigue, en la sección 2, presentamos el trabajo relacionado con el problema. Más adelante, en la sección 3, presentamos los conjuntos de datos y métodos utilizados en esta investigación. En la sección 4, presentamos el diseño del algoritmo. Después, en la sección 5, presentamos los resultados. Finalmente, en la sección 6, discutimos los resultados y proponemos algunas direcciones de trabajo futuras.

2. TRABAJOS RELACIONADOS

3.1 Manejo agroecológico de la roya del café

En los últimos cuatro años, millones de cafetaleros desde Perú hasta México han lidiado con brotes de la Roya del Café. Un alto número de asociaciones y cooperativas de caficultores, gobiernos locales y nacionales, técnicos, investigadores, programas y organismos internacionales han buscado respuestas rápidas frente a los impactos que trae consigo esta enfermedad directamente sobre las plantaciones, y sobre todo en las subsecuentes consecuencias que genera en la economía y seguridad alimentaria de un número significativo de familias, quienes ven en la producción y cosecha de café su principal fuente económica. En casi todos los casos, la respuesta fue la entrega de alimentos para paliar la inseguridad alimentaria, entrega de fungicidas, o crédito para comprar productos químicos que permitieran combatir los ataques de esta enfermedad en el campo.
.

3.2 La roya del cafeto en Colombia (impacto, manejo y costos del control)

En Colombia, tradicionalmente se han sembrado las variedades de café Típica, Borbón y Caturra, pertenecientes a la especie Coffea arábica, de excelente comportamiento agronómico pero susceptibles al hongo causante de la roya del cafeto, Hemileia vastatrix.

La roya del cafeto continúa siendo el principal problema patológico en el cultivo del café. Esta enfermedad está íntimamente ligada al desarrollo fisiológico del cultivo, al nivel de producción de la planta y a la distribución y cantidad de lluvia. A pesar de la información técnica generada y divulgada por Cenicafé, se ha encontrado que los caficultores no están controlando adecuadamente la enfermedad.

3.3 Prevención y control de la roya del café

La roya del café es la más severa enfermedad del cultivo desde que fue reportada en 1869. La enfermedad ha causado grandes pérdidas en la producción y en las áreas de cultivo en países de Asia, África y América. Una vez que la enfermedad aparece y se establece en un lugar no ha sido posible erradicarla, a pesar de múltiples estrategias implementadas por las familias productoras. En consecuencia, las familias han tenido que adaptarse y convivir con la roya; así, se han desarrollado prácticas culturales y diversos métodos de prevención y manejo.

3.4 La Roya del café: La enfermedad más limitante del cultivo

La Roya del café es la enfermedad más limitante del cultivo y cada año deja cientos de hectáreas de cafetales totalmente devastadas. Esta enfermedad es exclusiva de los cafetos y es causada por el hongo Hemileia vastatrix, que ataca específicamente a las hojas de las plantas de café. Cuando el ataque es severo, provoca la caída prematura de las hojas, dejando los árboles totalmente desnudos. Además, La planta al perder sus hojas no puede hacer el proceso de fotosíntesis y por tanto no hay la suficiente cantidad de savia disponible para el llenado de las cerezas.

3. MATERIALES Y MÉTODOS

En esta sección se explica cómo se recopilaron y procesaron los datos y, después, cómo se consideraron diferentes alternativas de solución para elegir un algoritmo de árbol de decisión.

3.1 Recopilación y procesamiento de datos

Obtuvimos datos del *Instituto Colombiano de Fomento de la Educación Superior* (ICFES), que están disponibles en línea en ftp.icfes.gov.co. Estos datos incluyen resultados anonimizados de Saber 11 y

Saber Pro. Se obtuvieron los resultados de Saber 11 de todos los graduados de escuelas secundarias colombianas, de 2008 a 2014, y los resultados de Saber Pro de todos los graduados de pregrados colombianos, de 2012 a 2018. Hubo 864.000 registros para Saber 11 y 430.000 para Saber Pro. Tanto Saber 11 cómo Saber Pro, incluyeron, no sólo las puntuaciones sino también datos socioeconómicos de los estudiantes, recogidos por el ICFES, antes de la prueba.

En el siguiente paso, ambos conjuntos de datos se fusionaron usando el identificador único asignado a cada estudiante. Por lo tanto, se creó un nuevo conjunto de datos que incluía a los estudiantes que hicieron ambos exámenes estandarizados. El tamaño de este nuevo conjunto de datos es de 212.010 estudiantes. Después, la variable predictora binaria se definió de la siguiente manera: ¿El puntaje del estudiante en el Saber Pro es mayor que el promedio nacional del período en que presentó el examen?

Se descubrió que los conjuntos de datos no estaban equilibrados. Había 95.741 estudiantes por encima de la media y 101.332 por debajo de la media. Realizamos un submuestreo para equilibrar el conjunto de datos en una proporción de 50%-50%. Después del submuestreo, el conjunto final de datos tenía 191.412 estudiantes.

Por último, para analizar la eficiencia y las tasas de aprendizaje de nuestra implementación, creamos al azar subconjuntos del conjunto de datos principal, como se muestra en la Tabla 1. Cada conjunto de datos se dividió en un 70% para entrenamiento y un 30% para validación. Los conjuntos de datos están disponibles en https://github.com/mauriciotoro/ST0245-Eafit/tree/master/proyecto/datasets .

	Conjunto de datos 1	Conjunto de datos 2	Conjunto de datos 3	Conjunto de datos 4	Conjunto de datos 5
Entrenamiento	15,000	45,000	75,000	105,000	135,000
Validación	5,000	15,000	25,000	35,000	45,000

Tabla 1. Número de estudiantes en cada conjunto de datos utilizados para el entrenamiento y la validación.

3.2 Alternativas de algoritmos de árbol de decisión

En lo que sigue, presentamos diferentes algoritmos usados para construir automáticamente un árbol de decisión binario. *(En este semestre, ejemplos de tales algoritmos son ID3, C4.5 y CART).*

3.2.1 CART

Los árboles de clasificación y regresión Cart son una alternativa al análisis tradicional de clasificación y/o discriminacion a la predicción tradicional (regresión). Entre las ventajas de estos árboles Cart podemos destacar su robustez a outliers, la invarianza en la estructura de sus árboles de clasificación o de regresión a transformaciones monótonas de las variables independientes, y sobre todo, su interpretabilidad.

3.2.2 ID3

El modelo de clasificación, también conocido como ID3, significa (inducción del árbol de decisiones), Un sistema de aprendizaje supervisado que utiliza una estrategia de clasificación "divide y vencerás", implementa métodos y técnicas para realizar procesos inteligentes que representan conocimiento y aprendizaje, automatizando así las tareas. Hoy en día es común escuchar que la palabra "inteligencia" o algunos de sus sinónimos se aplican a máquinas, sistemas, procesos e incluso productos del hogar que son ampliamente conocidos en todo el mundo. Su origen se debe a la aparición de la inteligencia artificial y los sistemas expertos.

3.2.3 C4.5

El algoritmo C4.5 genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente. El árbol se construye mediante la estrategia de profundidad-primero (depth-first). El algoritmo considera todas las pruebas posibles que pueden dividir el conjunto de datos y selecciona la prueba que resulta en la mayor ganancia de información. Para cada atributo discreto, se considera una prueba con resultados, siendo n el número de valores posibles que puede tomar el atributo. Para cada atributo continuo, se realiza una prueba binaria sobre cada uno de los valores que toma el atributo en

los datos. En cada nodo, el sistema debe decidir cuál prueba escoge para dividir los datos.

3.2.4 CLS

Un árbol de decisiones es un diagrama de los posibles resultados de una serie de decisiones relacionadas. Permite a las personas u organizaciones comparar posibles acciones en función de sus costos, probabilidades y beneficios. Se pueden utilizar para llevar a cabo una lluvia de ideas informal o para diseñar un algoritmo para predecir matemáticamente la mejor opción. Los árboles de decisión generalmente comienzan desde un solo nodo y luego se ramifican hacia posibles resultados. Cada uno de estos resultados crea otros nodos, que se pueden dividir en otras posibilidades. Esto le da una forma de árbol.

4. DISEÑO DE LOS ALGORITMOS

En lo que sigue, explicamos la estructura de los datos y los algoritmos utilizados en este trabajo. La implementación del algoritmo y la estructura de datos se encuentra disponible en Github[1].

4.1 Estructura de los datos

La estructura de datos que vamos a usar para este proyecto, es un árbol de decisión ID3, porque es capaz de tomar decisiones con gran precisión. . por ejemplo, nuestro tema principal en la predicción es diseñar un algoritmo basado en árboles de decisión, para predecir si un lote de café caturra está infectado con la roya o no; con ayuda de las principales variables que vamos a usar como lo son : iluminación, temperatura ambiental, humedad del suelo, humedad ambiental, temperatura del suelo, ph del suelo; tenemos no solo la información necesaria para tener una predicción completamente eficiente si no que tenemos muchos otros factores los cuales podemos utilizar en la predicción.

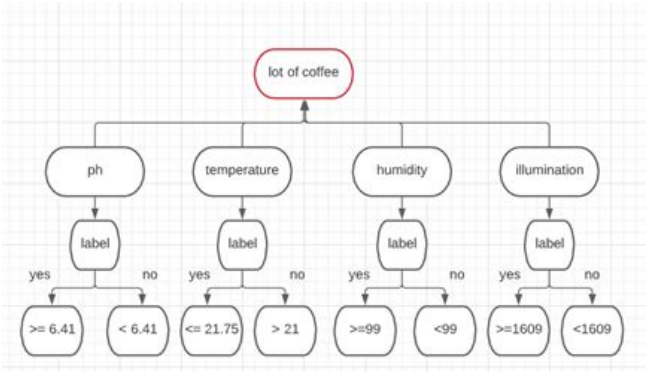

Figura 1: Ejemplo de el uso de los valores para la creación del árbol.

4.2 Algoritmos

Se testean varios intervalos en los valores del modelo y se selecciona el que tenga el valor gini ponderado más bajo, usando este método llegamos a que cuando la humedad del suelo es mayor a 67 entonces significa que esas plantas están aseguradas de no sufrir la enfermedad de roya (68), se divide el árbol y quedamos con un nuevo nodo de 304 valores para buscar el siguiente valor que pueda disminuir su impureza.

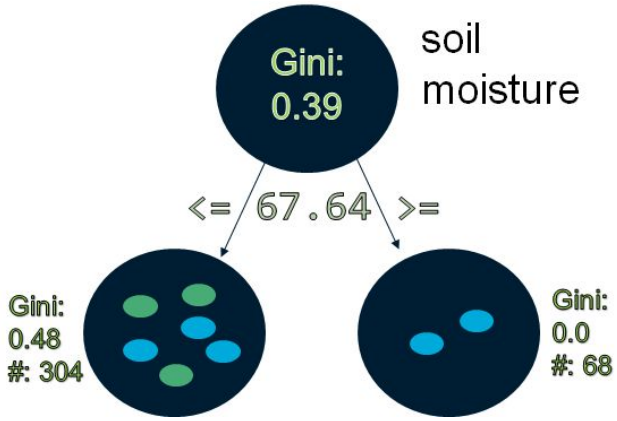

Figura 2: Una separación de nodo mostrando que valores se tomaron en cuenta.

4.2.1 Entrenamiento del modelo

El algoritmo va a probar cuál valor es el más efectivo para separar el árbol, y cuando lo separe va a tomar los hijos impuros y a repetir el proceso, va a guardar el orden de las instrucciones y así cuando reciba otro archivo de datos podrá recrearlas para predecir con precisión.

[1]

4.2.2 Algoritmo de prueba

Para probar que el prototipo del algoritmo está funcionando comparamos los valores del gini ponderado, el gini de los nodos hijos, el punto de inflexión de los hijos y la cantidad de valores con que cada hijo quedó. Primero probamos de entre los 6 valores el que tenía el gini más bajo era la humedad y que un nodo hijo tenía 0 impureza, esto puede suceder si el algoritmo le da 0 valores, pero fue diseñado para que no pasará, entonces este nodo hijo con 0 impureza es por que tiene 68 valores los cuales son iguales, esta decisión fue correcta por que en la primera inflexión se filtra 18% de los valores.

4.3 Análisis de la complejidad de los algoritmos

N es la cantidad de datos en el archivo.

M es el número de veces que se repite el ciclo de buscar el mejor número de división.

Algoritmo	La complejidad del tiempo
Entrenar el árbol de decisión	O(N*M)
Validar el árbol de decisión	O(N)
Reader.reader()	O(N)
Trainer.main()	O(N*M)
Trainer.trainer()	O(N*M)
Tree.main()	O(N)
Tree.tree()	O(N)

Se utilizan arrays de enteros para apuntar a los valores de cada nodo por lo cual es muy eficiente con la memoria, y el entrenador olvida los datos que no son eficientes, solo guardando los que son la mejor partición de cada nodo.

Algoritmo	La complejidad de memoria
Entrenar el árbol de decisión	O(N*M)
Validar el árbol de decisión	O(N)

4.4 Criterios de diseño del algoritmo

El algoritmo fue elegido por su simpleza y flexibilidad, para poder tomar en cuenta tantos valores distintos y con tantas implicaciones la mejor opción es tomar un acercamiento eficiente. En vez de dividir e ir moviendo de lugar todos los datos cada vez que se hace una división del nodo, preferimos que cada nodo sólo guardará un puntero a los datos que posee y otra información importante propia de cada nodo, esto fue para ahorrar mucho trabajo y memoria que se gastaría moviendo los datos.

5. RESULTADOS

5.1 Evaluación del modelo

Exactitud: 227 estudiantes fueron identificados correctamente por el modelo de 300, lo que lleva a 76% probabilidad de acierto.

Precisión: 148 plantas fueron diagnosticadas con roya, de las cuales 83 realmente tenían la enfermedad.

Sensibilidad: El modelo identificó a 168 plantas con roya, y había 181 plantas con la enfermedad.

5.1 Evaluación del modelo en entrenamiento

	Conjunto de datos 1	*Conjunto de datos 2*	*Conjunto de datos 3*
Exactitud	85%	81%	62%
Precisión	90%	66%	52%
Sensibilidad	96%	80%	84%

5.2 Tiempos de ejecución

Tiempo requerido para correr el proceso mil veces:

	Conjunto de datos 1	*Conjunto de datos 2*	*Conjunto de datos 3*
Tiempo de entrenamiento	12.48 s	17.57 s	24.6 s
Tiempo de validación	0.00099 s	0.0019 s	0.0019 s

5.3 Consumo de memoria

	Conjunto de datos 1	*Conjunto de datos 2*	*Conjunto de datos 3*
Consumo de memoria	9 KB	11 KB	25 KB
Entrenar	90	114	127
Validar	24	40	227

6. DISCUSIÓN DE LOS RESULTADOS

Explique los resultados obtenidos. ¿Son la precisión, exactitud y sensibilidad apropiadas para este problema? ¿El modelo está preajustado? ¿Es el consumo de memoria y el consumo de tiempo sib apropiados? *(En este semestre, de acuerdo con los resultados, ¿se puede aplicar esto para dar becas o para ayudar a los estudiantes con baja probabilidad de éxito? ¿Para qué es mejor?)*

6.1 Trabajos futuros

Nos interesaría investigar alternativas de librerías externas que podrían aumentar el alcance y flexibilidad del código.

AGRADECIMIENTOS

Esta investigación fue apoyada parcialmente por los monitores de la materia de estructuras de datos y algoritmos 1 de la universidad Eafit , reconociendo a los estudiantes Simón Marín Giraldo, Daniel Alejandro Mesa y Miguel Correa, como mayor apoyo en el transcurso de la realización de este informe . la metodología y los temas tratados en este informe, nos permitió tener cierta libertad en el manejo de los temas vistos en la materia, permitiéndonos así , una profundización más amplia sobre los temas tratados en clase..

REFERENCIAS

La referencias se hacen con el formato de referencias de la ACM. Lea las directrices de ACM en http://bit.ly/2pZnE5g

A modo de ejemplo, consideremos estas dos referencias:

1.*Adobe Acrobat Reader 7, Asegúrate de que el texto de las secciones de referencia es está alíneado a la derecha y no justificado. http://www.adobe.com/products/acrobat/.*

2. Fischer, G. y Nakakoji, K. Amplificando la creatividad de los diseñadores con entornos de diseño orientados al dominio. en Dartnall, T. ed. Artificial Intelligence and Creativity: An Interdisciplinary Approach, Kluwer Academic Publishers, Dordrecht, 1994, 343-364.

3.1 Timarán-Pereira, R., Caicedo-Zambrano, J., & Hidalgo-Troya, A. (2019). Árboles de decisión para predecir factores asociados al desempeño académico de estudiantes de bachillerato en las pruebas Saber 11°. REVISTA DE INVESTIGACIÓN, DESARROLLO E INNOVACIÓN, 9(2), 363-378.

3.2 Gabalán-Coello, Jesús; Vásquez-Rizo, Fredy Eduardo SABER 11 y rendimiento universitario: un análisis del progreso en el plan de estudios Ciencia, Docencia y Tecnología, vol. 27, núm. 53, noviembre, 2016, pp. 135-161 Universidad Nacional de Entre Ríos Concepción del Uruguay, Argentina

3.3 Marly Tatiana Celis, Óscar Andrés Jiménez y Juan Felipe Jaramillo Maestría en Economía, Universidad de Manizales