ÁRBOL DE DECISIÓN PARA EL MONITORIEO Y PREDICCIÓN DEL HONGO DE LA ROYA EN LA PLATA DEL CAFÉ

Antoine Chavane de Dalmassy C.
Universidad Eafit
Colombia
achavaned@eafit.edu.co

Santiago Ospina Idrobo
Universidad Eafit
Colombia
sospinai@eafit.edu.co

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

RESUMEN:

El objetivo de este proyecto es analizar y buscar la solución más adecuada y eficiente para un problemática muy importante acerca de la plaga de la roya café que este es el principal problema fitosanitario en el café, esta es una problemática que requiere de una muy buena solución ya que el café es la principal exportación agrícola en el país, donde este se da por culpa de diagnósticos que se realizan muy tarde y al hacer el diagnostico luego de un largo tiempo se verán altas perdidas en el cultivo del café.

La solución a este problema es muy importante por tres sencillas razones: La primera es que Colombia es el tercer país en el mundo por debajo de Brasil y Vietnam en la lista de los principales productores de café en el mundo; La segunda es que el café es la principal exportación agrícola en Colombia y como ultimo hasta el momento se sabes que hay 563.000 familias que dependen de este cultivo. Por lo que hallando la solución a este problema podremos mejorar y aprovechar más a fondo este cultivo que es una de las cosas más importantes en nuestro país, así que trataremos de encontrar la solución y forma óptima para solucionar este problema.

Palabras claves:

Agricultura, id3, arboles de decisiones, Cart, data, café, roya del café.

Palabras claves de la ACM:

CCS -> Computing methodologies -> Modeling and simulation -> Simulation evaluation

CCS -> Information systems -> Data management systems -> Data structures -> Data access methods

1. INTRODUCCIÓN:

Como ya sabemos el cultivo que mas produce y exporta Colombia a nivel de lo agrícola es el cultivo de café, por lo que cada vez más los productores de café buscan la forma de mejorar cada vez más el producto, la forma en la que se ve y se vende el producto.

En este documento hablaremos acerca de la problemática de la roya en cultivos de café y como podremos monitorearla y buscarle una solución; la roya de café es una enfermedad que es causada por un hongo llamado Hemileia vastatrix y el café es el único portador conocido de este hongo, esta infección en los cultivos de café llego a Brasil a mediados de los 70´s, a Centroamérica aproximadamente en el año 1976 y a Colombia en los años 80´s.

2. PROBLEMA*:*

El problema acerca de las royas de café a sido considerado una de las enfermedades más catastróficas en la historia, este problema se encuentra entre las siete pestes que han dejado mas perdidas en los últimos 100 años, este problema debe ser tratado con delicadeza ya que este si no tiene un buen trato llegaría a generar un impacto socio económico que este podría llegar a generar la roya de café es de una dimensión tan grande que podría llegar a ser incalculables.

Por lo que es de mucha delicadeza tener una buena solución con la cual tratar esta enfermedad en los cultivos de café, ya que, si llega a tener éxito y se disminuye la epidemia de esta plaga, la economía se disparara y se obtendrán mejores ingresos por los lados de la agricultura y de uno de los elementos mas exportados y producidos en Colombia.

3. TRABAJOS RELACIONADOS:

3.1 Algoritmo ID3

Este algoritmo es utilizado para el uso de la inteligencia artificial, el uso de este algoritmo consiste en la búsqueda de hipótesis dado un conjunto de cosas, estas cosas son una serie de tuplas de valores donde estos son considerados y/o denominados atributos, donde el atributo a clasificar es el objetivo el cual es de tipo binario ósea: positivo o negativo, si o no, verdadero o falso.

El algoritmo ID3 realiza su trabajo mediante un árbol de decisión; un árbol de decisión es un modelo en el cual dado un conjunto de datos se fabrican diagramas de construcciones lógica, los cuales sirven para representar y categorizar una serie de condiciones que ocurren de forma sucesiva, y llevan a la solución de un problema, los elementos de un árbol de decisión son:

Nodos: Los cuales contendrán atributos.

Arcos: Los cuales contienen valores posibles de un nodo padre.

Hojas: son las que se encargan de clasificar el nodo como positivo o negativo.

Un problema que se pueda solucionar con este algoritmo seria:

Una persona tiene como tipo de sangre: O+, y necesita que alguien le done sangre para una operación, así que hay tres de sus familiares que se prestan para esta decisión, pero no se sabe cual es el donante correcto para este tipo de sangre, los grupos sanguíneos de los familiares son los siguientes:

Primer familiar: A+

Segundo familiar: B-

Tercer familiar: O-

Lo que hace el algoritmo en este caso es lo siguiente:

Si el grupo sanguíneo del receptor es O+..

Si es diferente de O(+)(-) devolver falso,

Si es Igual que O(+)(-) devolver verdadero,

El Resultado sería:

Primer familiar: Falso.

Segundo familiar: Falso.

Tercer familiar: Verdadero.

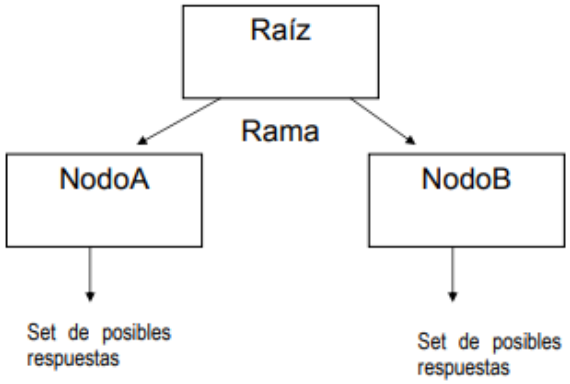
3.2 Algoritmo C4.5

El algoritmo C4.5 fue desarrollado EN 1993 para ser una extensión (mejora) del algoritmo ID3 que se desarrolló en 1986. Éste genera un árbol de decisión a partir de los datos mediante particiones realizadas recursivamente.

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 n 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.

El C4.5 forma parte de la familia de los TDIDT (Top Down Induction Trees), junto con antecesor el ID3. El C4.5 se basa en el ID3, por lo tanto, la estructura principal de ambos métodos es la misma. El C4.5 construye un árbol de decisión mediante el algoritmo "divide y vencerás" y evalúa la información en cada caso utilizando los criterios de Entropía, Ganancia o proporción de ganancia, según sea el caso.

Están formados por:
- Nodos: Nombres o identificadores de los atributos.
- Ramas: Posibles valores del atributo asociado al nodo.
- Hojas: Conjuntos ya clasificados de ejemplos y etiquetados con el nombre de una clase.



3.3 Modelo de árbol de decisión CART

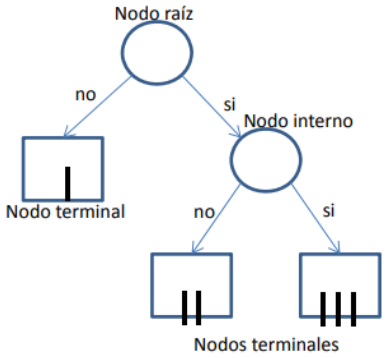
A Los árboles de clasificación y regresión (CART=Classification and Regression Trees) son una alternativa al análisis tradicional de clasificación/discriminación o a la predicción tradicional (regresión). Entre las ventajas de estos árboles CART podemos destacar su robustez a outliers, la invarianza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

Son árboles de regresión cuando la variable dependiente es continua y árboles de clasificación cuando la variable dependiente es de tipo cualitativo. En esencia, se trata de dar con un esquema de múltiples dicotomías o bifurcaciones, anidadas en forma de árbol, de manera que siguiendo cada una de las ramas del árbol obtengamos, al final, una predicción para la clase de pertenencia (clasificación) o para el valor que toman (regresión) los individuos que cumplen con las propiedades que se han ido exigiendo en las distintas bifurcaciones.

El proceso puede esquematizarse en 4 fases: construcción (building) del árbol, parada (stopping) del proceso de crecimiento del árbol (se constituye un árbol máximo que sobreajusta la información contenida en nuestra base de datos), podado (pruning) del árbol haciéndolo más sencillo y dejando sólo los nodos más importantes y, por último, selección (selection) del árbol óptimo con capacidad de generalización.

La construcción del árbol comienza en el nodo raíz, que incluye todos los registros de la base de datos. A partir de este nodo el programa debe buscar la variable más adecuada para partirlo en 2 nodos hijos. Para elegir la mejor variable debe utilizarse una medida de pureza (purity) en la valoración de los 2 nodos hijos posibles (la variable que consigue una mayor pureza se convierte en la utilizada en primer lugar, y así sucesivamente). Debe buscarse una función de partición (splitting function) que asegure que la pureza en los nodos hijos sea la máxima. Una de las

funciones más utilizada es la denominada Gini (se alcanza un índice de pureza que se considera como máximo).

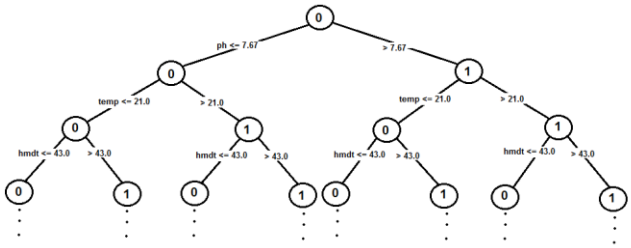


3.4 CHAID

El problema para el cual se creó este algoritmo radica en el turismo y saber cuántas personas irán a un destino en un tiempo determinado, como, por ejemplo: En la ciudad de Medellín cuando se celebra la feria de las flores, ¿cuánta cantidad de turistas llega en esas fechas aproximadamente? El algoritmo CHAID se encarga de comprar un tipo de información acerca de la segmentación del mercado turístico, en conclusión, este algoritmo hace un estudio y/o análisis en el cual calcula que destino será el más visitado por turistas en un momento determinado.

4. CART

El árbol de decisiones CART se usa para abreviar y poder usarse para problemas de modelación predictiva de clasificación o regresión.



4.2 Analisis de complejidades

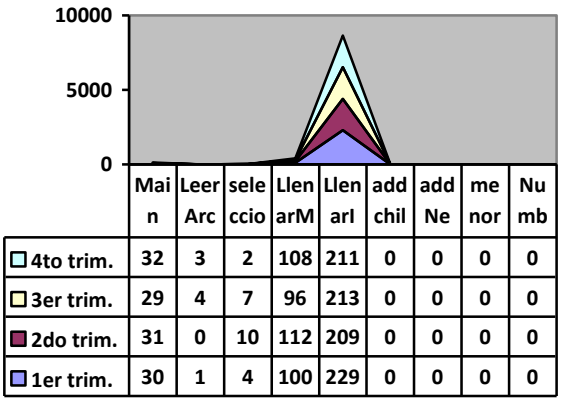
Main	O (n x m)
LeerArchivo	O (n x m)
Seleccionar	O (n)
LlenarImpureza	O (n x m)
LlenarMatriz	O (n x m)
addChild	O (1)
addNewNode	O (1)
menores	O (n)
NumberOfNodesInTree	O (n)

4.3 Criterios de diseño de estructura de datos:

Elegimos este diseño de estructura de datos ya que es el mas optimo y eficiente a la hora de manipular los datos y poder crear en árbol con respecto a sus nodos izquierdos y derechos, con esta estructura de datos buscamos la menor complejidad posible y el menor gasta de tiempo y de memoria a la hora de ser ejecutado el programa.

4.4 Resultados obtenidos en tiempo (ms):

Main	30
LeerArchivo	1.0
seleccionar	5.0
LlenarImpureza	2000
LlenarMatriz	100
addChild	0.0
addNewNode	0.0
menores	0.0
NumberOfNodesInTree	0.0



4.5 Resultados obtenidos en memoria (4 archivos)

data_set.csv	7.64 MB
data_set_balanced.csv	6.98 MB
data_set_train.csv	6.84 MB
Data_set_test.csv	6.56 MB

5. Conclusiones

Lo mas importante de esta estructura de datos es que gracias a esta se puede identificar a partir de unos datos enviados por sensores colocados estratégicamente en un cultivo de café se puede deducir si un cultivo de estos esta infectad de roya y así saber si se puede fumigar o no.

Entre los resultados mas importantes esta saber que cultivo de café esta infectado y saber cuál no.

Entre la posibilidad del futuro de este trabajo esta poder analizar más rápido los datos con un menor uso de memoria y por automatizar todo el cultivo lo cual conlleva a manejar mas variables y sensores.

REFERENCIAS

1. López, B. ~ *ALGORITMO C4.5* ~. Nuevo Laredo, Tamaulipas, noviembre del 2005.

2. Trujillano, J. Sarria-Santamera, A. Esquerda, A. Badia, M. Palma, M. and March, J. *Aproximación a la metodología basada en árboles de decisión (CART). Mortalidad hospitalaria del infarto agudo de miocardio*. Gac Sanit vol.22 no.1 Barcelona ene./feb. 2008.

3. *Journal of Destination Marketing & Management Volume 5, Issue 3, September 2016, Pages 275-282:*

https://www.sciencedirect.com/science/article/pii/S2212571X1600007X

4.https://es.wikipedia.org/wiki/Algoritmo_ID3