IMPLEMENTACIÓN DE ÁRBOLES DE DECISIÓN EN LA PREDICCIÓN DEL ÉXITO ACADÉMICO

Stiven Yepes Vanegas
Universidad Eafit
Colombia
esyepesv@eafit.edu.co

Sara Rodríguez V
Universidad Eafit
Colombia
srodriguev@eafit.edu.co

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

1. INTRODUCCIÓN

En la actualidad se cuenta con grandes bases de datos en las instituciones, como lo es la base de datos de la universidad EAFIT, sin embargo, éstos son poco útiles si solo se almacenan y no se les relaciona entre sí para encaminarse a responder interrogantes más apremiantes, como lo es la probabilidad de éxito académico, problemática que se planea abordar a través de la creación de árboles de decisión.

2. PROBLEMA

El problema del éxito académico cuenta con muchas variables sociales, económicas, políticas, entre otras, pero a través de la tecnología, y en especial del aprendizaje supervisado aplicado en árboles de decisión, se puede buscar de manera más directa una relación entre los datos recogidos por el ICFES y la probabilidad de éxito académico de un estudiante de la Universidad EAFIT para hacer predicciones útiles que permitan crear medidas preventivas y de acompañamiento.

3. TRABAJOS RELACIONADOS

3.1 ID3

Existen diferentes métodos para construir un árbol de decisión, entre ellos está el algoritmo ID3, desarrollado por J. Ross Quinlan en 1983, cuyo nombre hace referencia a *Induction Decision Tree*. Hace parte de los algoritmos TDIDT *(Top-Down Induction of Decision Trees*).

Este algoritmo se basa principalmente en la entropía o probabilidad de ocurrencia de un evento, la cual es utilizada para dividir un conjunto de datos en subgrupos más pequeños de forma recursiva. Sus principales componentes son los nodos, que identifican los atributos; las ramas, que son los posibles valores con relación al nodo y las hojas, que son los conjuntos ya clasificados.

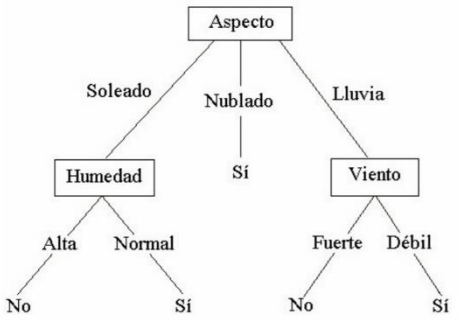

Figura 1: Ejemplo de un árbol con ID3.
Ver referencia 1.

En el ejemplo anterior vemos se subdividen los atributos de la mejor manera, esto lo hace siguiendo la siguiente secuencia: primero calcula la entropía de cada atributo, después divide el conjunto original en subconjuntos con menor entropía y crea nodos de estos, hace esto varias veces recursivamente y el resultado final son las hojas con las etiquetas o clasificaciones del problema en cuestión.

Algunos inconvenientes que tiene este algoritmo son el favoritismo por aquellos atributos que no necesariamente son los más útiles, la generación de grandes arboles y que solo es aplicable a problemas de clasificación y diagnóstico.

3.2 C4.5

El algoritmo C4.5 es una mejora hecha por J. Ross Quinlan al algoritmo ID3 en 1993, y como este hace parte de la familia de los algoritmos TDIDT. Al igual que su antecesor, Este algoritmo tiene los mismos componentes del ID3 y hace uso de la entropía, pero ahora acompañada de la estrategia de profundidad-primero (*depth-first*). La gran similitud que hay entre estos dos algoritmos se puede ver en el siguiente gráfico:

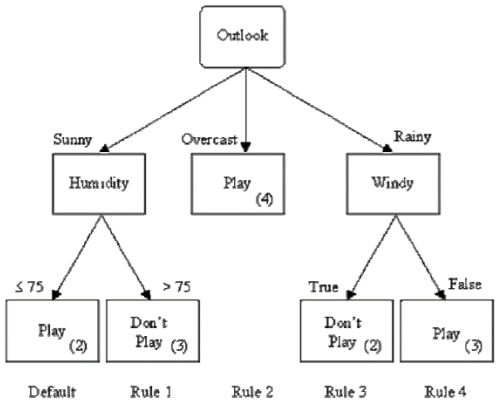

Figura 2: Ejemplo de un árbol con C4.5.
Ver referencia 2

Aquí se puede apreciar como esta versión mejorada del algoritmo no solo considera valores discretos, sino que también considera valores continuos. Para los atributos discretos se considera una prueba con todos los posibles

valores que puede tomar el atributo y para los atributos continuos se realizaba una prueba binaria sobre estos datos.

3.3 CART

El algoritmo CART extrae su nombre *de Classification and Regression Trees*, término acuñado por Leo Breiman, es de gran relevancia y sirve de base para algoritmos posteriores como *Random Forest* o *Bagged Decision Tree*.

Se representa con un árbol binario, donde cada raíz tiene una variable (x), que se divide en dos ramas a través de una sentencia, normalmente numérica, a la que se le asigna un valor booleano (*True or False*) cada rama posteriormente se divide de la misma forma. Las hojas contienen una variable (y) que es usada para hacer una predicción.

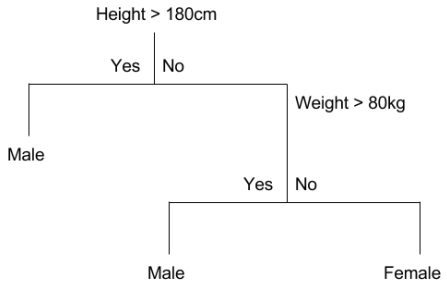

Figura 3: Ejemplo hipotético de árbol usando CART.

Ver referencia 3.

En el ejemplo descrito en la gráfica se evidencia la naturaleza booleana de los atributos de un árbol CART, y se evidencia la importancia de hacer una buena elección de variables para evitar generalizaciones erróneas, como que una persona que mida más de 180 cm siempre será hombre.

Para construir un modelo CART se debe seleccionar y adecuar los datos de entrada y separar los resultados de esas variables hasta que se cree un árbol adecuado. La selección de que variable usar y el punto de corte o separación se realiza a través de un *Greedy Algorithm* para minimizar el costo de correr el árbol.

3.4 CHAID

El acrónimo CHAID se refiere a *Chi-squared Automatic Interaction Detector*. Es uno de los métodos de clasificación propuesto por Kass, y se considera un descendiente de THAID. Este algoritmo construye árboles no binarios, a través del uso de un algoritmo bastante útil para el análisis de *data sets* de gran tamaño llamado Chi Square. El hecho de que trabaje con tablas de frecuencia multidireccional lo vuelve muy popular en el ámbito del mercadeo, en especial de estudios de segmentación de mercado.

El algoritmo crea vaticinadores categóricos dividiendo distribuciones continuas en un conjunto de categorías. Entonces el algoritmo itera los vaticinadores realizando pruebas, mezclando categorías y ajustando valores, escogiendo las ramas más significativas hasta el momento que no se puedan realizar más divisiones.

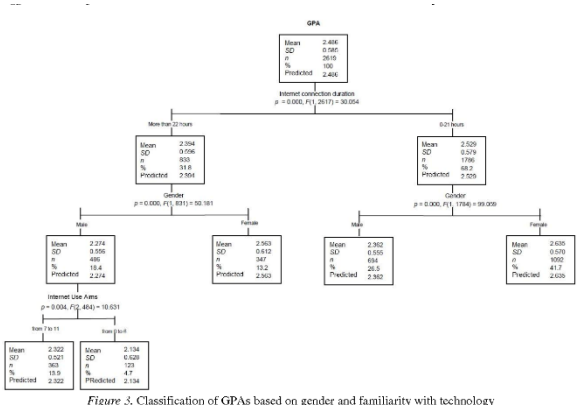

Figure 3. Classification of GPAs based on gender and familiarity with technology

Figura 4: Ejemplo de árbol usando CHAID.

Ver referencia 5.

En el anterior ejemplo se evidencia un árbol CHAID en acción en un ambiente similar al que se efectuará en el proyecto en cuestión, analizando hábitos de estudiantes universitarios. Se aprecia que hay más variables involucradas que en el algoritmo analizado anteriormente, se ve las probabilidades estadísticas asociadas al vaticinador de cada rama, junto con su desviación.

Un principal inconveniente es el tamaño que los árboles pueden llegar a tener, no por un motivo computacional, pero de interpretación humana posterior de los árboles creados. También que necesita un volumen considerable de datos para funcionar correctamente.

REFERENCIAS

1. López Takeyas, B (2015) Algoritmo ID3. recuperado en 07/02/2020 de: http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/IA/ID3.pdf

2. Espino J, Tijerina J, Cedano M, Amaya E, Pérez J, Chiñas A. (2015) Inteligencia Artificial: Algoritmo C4.5. recuperado en 08/02/2020 de: http://www.itnuevolaredo.edu.mx/takeyas/Apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/C4.5/C4.5(2005-II-B).pdf

3. Brownlee J. (2016) Classification And Regression Trees for Machine Learning. Machine Learning Mastery. Recuperado en 05/02/2020 de: https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

4. StatSoft Inc (2013) Electronic Statistics Textbook. Tulsa. OK: StatSoft. Recuperado en 05/02/2020 de: http://www.statsoft.com/textbook/CHAID-Analysis

5. B. Bahar, K. Eylem (2015) Applying The CHAID Algorithm to Analyze How Achievement is Influenced by University Students' Demographics, Study Habits, and Technology. Recuperado en 06/02/2020 de: https://www.semanticscholar.org/paper/Applying-The-CHAID-Algorithm-to-Analyze-How-is-by-Baran-Kili%C3%A7/eafaa5f8c88192a4c586e5bb045c5c0e049980ed