ARBOLES DE DECISIÓN PARA LA EVALUACION DE RESULTADOS EN LA EDUCACIÓN SUPERIOR

David Vergara Patiño
Universidad Eafit
Colombia
dvergarap@eafit.edu.co

Andrés Gómez Arango
Universidad Eafit
Colombia
mflorezr@eafit.edu.co

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

RESUMEN:

El problema que vamos a tratar es que vamos a medir el éxito académico de los estudiantes de educación superior por medio de árboles de decisión, con esto evaluaremos el éxito académico de los estudiantes y podremos saber que variables son las que más afectan, este éxito académico. Este problema es importante porque nos puede servir de medida, para valorar la educación superior en el país, y también identificar cuales son las mayores falencias de los estudiantes en temas académicos. Algunos problemas relacionados son por medio de redes neuronales, o máquinas de vectores de soporte. Entre otros.

1. INTRODUCCIÓN:

Dia a día, la deserción en las universidades, en conjunto con las pocas posibilidades para acceder a las universidades y con la dificultad para conseguir empleo, están llevando a que los estudiantes se desmotiven cada vez más, esto se traduce en problemas académicos, debido a que ya los estudiantes no le ven objeto a esforzarse, y al final en la prueba saber pro, a los estudiantes les va cada vez peor. Históricamente el problema ha ido aumentando considerablemente, cada vez a las personas les importa menos los exámenes, aquí también se ve afectadas las empresas, debido a que cada vez los servicios de educación son de menor calidad, según cifras solo el 21,8% de las instituciones de educación superior cuenta con acreditación de alta calidad:

2. PROBLEMA

El problema al que nos enfrentamos es que por medio de arboles de decisión, queremos estimar la probabilidad de que una persona obtenga más del promedio, resolver este problema, nos ayudara a entender cuales son las variables que mas afectan los puntajes de las saber pro, con esto se podrá hacer un informe que le podría ayudar a las universidades a tratar esas falencias, además de eso nos ayudara a ver la calidad de la educación superior en Colombia.

3. TRABAJOS RELACIONADOS

3.1. ID3

El ID3 construye un árbol de decisión de manera directa de arriba hacia abajo, sin usar backtracking, y basándose solamente en los ejemplos iniciales proporcionados. Y usa el concepto de ganancia de información n para así tomar el atributo mas útil. En resumen, escoge las mejores preguntas, las que pueden tener mayor ganancia de información. Esto lo hace utilizando la incertidumbre de Shannon, la cual mide la incertidumbre de una muestra.

Entradas:

Las entradas son un conjunto de ejemplos, descritos en una serie de atributo-valor. Es una tabla donde la fila es el ejemplo y en las columnas se almacenan los valores de cada atributo. Y uno de esos atributos debe ser el objetivo de la predicción(clase).

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

Figura 1: ejemplo de entrada.

La salida será un árbol que separa los ejemplos de acuerdo con las clases a las que pertenece.

Condiciones:

Clases predefinida: todo ejemplo tiene un valor de clase definido.

Clases discretas: siempre se conoce cuantas clases puede haber.

3.2. C4.5

El algoritmo C4.5 es una mejora del algoritmo ID3, en el que se mejoran que se pueden agregar atributos, discretos y continuos, falta de valores y además que se pueden poder los arboles después de su construcción, optimizando los valores arrojados. En el resto de las cosas es igual al algoritmo ID3, solo que más eficiente.

Atributos discretos y continuos: en este caso se puede agregar por ejemplo variables cualitativas, en la que 1 es si y 0 es no.

Falta de valores: en los ejemplos de entrada pueden faltar valores, y cuando esto pase se agregar un ? y lo que hace es que lo estima la todos los datos.

Corte de los árboles: esto lo que busca es construir arboles más pequeño.

3.3. C5

De todos los algoritmos de arboles de decisión, el C5 es uno de los mas usados, debido a que es muy preciso tanto, como para compararlo con redes neuronales, y maquina vector soporte. Y con la diferencia de que es más fácil de implementar y de comprender.

Poda de los árboles: un beneficio de este algoritmo a diferencia de ID3 y de C4.5, es que su poda es muy precisa, el algoritmo después de hacer todo el árbol completo elimina los nodos, que no tienen importancia automáticamente.

Utiliza el limite binomial de confianza para reducir el tamaño de los árboles, con esto lo que hace es que elimina las ramas que no afectan la confianza del modelo.

3.4. CART

El algoritmo CART, es un algoritmo de clasificación, se diferencia de los anteriormente vistos ya que, los anteriores como parámetro de jerarquización utilizaban la entropía, este por el contrario utiliza el índice GINI, con este lo que hace es mostrar que tan importante es un nodo en el árbol.

Ahora para la poda de los árboles, este método usa la poda de complejidad de costos, este lo que busca es que el árbol no pierda efectividad en la predicción ya que se puede dar que al final el árbol arroje una información basada en muy pocos datos que puede terminar afectando la predicción.

El CART es un algoritmo muy simple y utilizado debido a que se puede implementar variables categóricas como cuantitativas, además porque conduce a un modelo muy simple para explicar porque las observaciones se clasifican en un determinado grupo, ayudando mucho al análisis de las predicciones.

4. REFERENCIAS

1. Fernando Sancho Caparrini**.** Aprendizaje Inductivo: Árboles de Decisión. Retrieved February 8, 2020. http://www.cs.us.es/~fsancho/?e=104

2. Octavia. Decisión tres-C4.5.Retrieved February 8,2020. https://octaviansima.wordpress.com/2011/03/25/decision-trees-c4-5/

3. Ricardo Fraiman. Clasificación Supervisada. Métodos jerárquicos CART. Retrieved February 8,2020. http://www.pedeciba.edu.uy/bioinformatica/datamining/curso_Data_Mining_clase_6.pdf

4. Czar Yober. Determining Creditworthiness for Loan Applications Using C5.0 Decision Trees. Retireved February 8,2020. https://rpubs.com/cyobero/C50