Árboles de Decisión para el Análisis Comparativo de Medidas de Evaluación de Nodos

Fernando Enrique Acevedo, Simón Figueroa, Claudia Ortiz, Juan Cruz Soto
Grupo n° 9 Inteligencia Artificial,
Universidad Tecnológica Nacional
Resistencia (Chaco), Argentina
{facevedo326, simonfigctes, claudiaortiz1311, juancruzzsoto}@gmail.com

Resumen. Pueden existir múltiples caminos obtenidos a la hora de implementar árboles de decisión asociados a la solución de un problema, lo cual se puede traducir en un árbol más grande, más ambiguo y por lo tanto más difícil de entender para los humanos. Normalmente son sometidos a datos con decenas de miles de transacciones y es por eso que nos interesa estudiar cómo se comporta y cómo afecta a los resultados de la implementación, pero por sobre todo, saber elegir cuál modelo se adapta mejor al conjunto de datos con el cuál estemos trabajando. En este informe se presentará dos modelos de construcción de Árboles de Decisión aplicando a sus nodos, la Ganancia de Información y la Tasa de Ganancia, se evaluará en qué contexto y bajo qué condiciones es conveniente usar uno por sobre el otro y para este proceso se usarán distintas métricas, que son la Puntería, Precisión, Exhaustividad y FScore.

Palabras clave: Ganancia de información, Tasa de ganancia, Conjunto de datos, Puntería, Precisión, Exhaustividad, FScore, Árboles de Decisión, Variabilidad, Atributos, Clase.

1. INTRODUCCIÓN

La automatización de procesos y el estudio de datos cada vez toma un rol más importante en ámbitos industriales, de negocios, entre otros; por lo tanto, la existencia de algoritmos capaces de tomar decisiones o llevar a cabo análisis de datos es vital, es por ello que es una prioridad que estos algoritmos sean lo más precisos posible, y en caso de que fallen, tener conocimiento de la probabilidad de ocurrencia de errores.

El proceso de toma de decisiones puede resultar complejo y a veces el riesgo de equivocarse es muy alto, con lo cuál debe ser lo más objetivo y preciso posible. Es por ello, que una técnica interesante es la implementación de árboles de decisión, que a través de un algoritmo (propuesto por Ross Quinlan [1]) llamado C4.5 permite determinar cuáles son los atributos que mayor información entregan del conjunto,

tomando como referencia los valores de un atributo especial llamado *clase* que representa las diferentes elecciones que se tomarán. Este informe presenta la implementación de dicho algoritmo y además busca comparar (a través de distintas simulaciones) cuál de los dos criterios de selección de nodos a explotar (ganancia de información y tasa de ganancia de información) son más eficientes para la toma de decisiones correctas.

El estudio se llevó a cabo a través de un programa desarrollado en Python que opera con un archivo en formato CSV en donde se tiene cargada una tabla con valores para diferentes atributos y una clase.

2. MARCO TEÓRICO

2.1. CONCEPTOS BÁSICOS

Un conjunto de datos utilizado en la tarea de aprendizaje consta de un conjunto de registros de datos, que se describen mediante un conjunto de atributos A donde | A | denota el número de atributos o el tamaño del conjunto A. El conjunto de datos también tiene un atributo objetivo especial C, que se denomina atributo de clase. Esto a su vez constituye un conjunto de datos para el aprendizaje que es simplemente una tabla relacional. Cada registro de datos describe una experiencia pasada.

Dado un conjunto de datos D, el objetivo del aprendizaje es producir una función de clasificación / predicción para relacionar valores de atributos en A y clases en C.

La precisión de un modelo de clasificación en un conjunto de pruebas se define como:

$$Puntería = \frac{Número\ de\ clasificaciones\ correctas}{Número\ total\ de\ casos\ de\ prueba} \quad \textbf{(1)}$$

donde una clasificación correcta significa que el modelo aprendido predice la misma clase que la clase original del caso de prueba.

En los casos en que el conjunto de datos de prueba posee una clase cuya distribución de valores es muy sesgada (la cantidad de valores comparada con otra es mucha) la Puntería deja de ser una medida que representa que tan confiable es el modelo para clasificar, y esto ocurre porque aquellos valores de clase con muy poca cantidad representa un costo muy grande si su clasificación es errónea. Por ejemplo, si se tienen 100 clasificaciones, dónde 95 corresponden al valor “Riesgo Bajo” y 5 a “Riesgo Alto” de la clase “Riesgo de Inversión” y además la Puntería calculada es 0.98 (falló en clasificar un Riesgo Bajo y un Riesgo Alto), el costo de clasificar mal una inversión como Riesgo Alto cuando no corresponde no es grave, pero en caso contrario, si se etiqueta como Riesgo Bajo a una inversión que no corresponde, el costo de ello es muy alto. Es por esto que se utilizan nuevas medidas capaz de

comprender estos casos, estas son la Precisión, Exhaustividad y FScore, provenientes de la confección de la Matriz de Confusión [2][4].

Se selecciona a un valor de clase que deseamos tomar interés (recomendablemente el valor cuya ocurrencia es la más baja) y se la denomina *valor Positivo*, la suma de las ocurrencias de los otros valores de la clase corresponderá al *valor Negativo*. Para poder calcular las medidas antes mencionadas es necesario obtener la matriz de confusión ya que esta contiene información acerca de los resultados actuales y predichos por el clasificador [2].

Tabla 1. Composición de la Matriz de Confusión [4].

	Negativo Clasificado	**Positivo Clasificado**
Negativo Actual	VN	FP
Positivo Actual	FN	VP

Donde:

VN: Verdadero Negativo; FN: Falso Negativo; FP: Falso Positivo; VP: Verdadero Positivo.

Precisión [2]: medirá el número de positivos clasificados correctamente dividido por el total de números de ejemplo que son clasificados como positivos:

$$p = \frac{VP}{VP+FP} \quad \textbf{(2)}$$

Es una excelente medida para determinar la confiabilidad del modelo cuando el costo de FP es muy alto [4].

Exhaustividad: es el número de VP clasificados dividido por el número total de positivos actuales en el set de pruebas.

$$r = \frac{VP}{VP+FN} \quad \textbf{(3)}$$

Es usado, a diferencia de la Precisión, cuando el alto costo está asociado a los FN [4].

FScore: es la media armónica de las medidas antes mencionada, esta media tiende a ser más cercana a la menor de las medidas que calcula, y esto se debe a que la principal propiedad de la media armónica es que se ve muy poco influenciada por valores grandes, pero si por valores más pequeños.

$$F = 2\frac{pr}{p+r} \quad \textbf{(4)}$$

Es la mejor medida para usar en casos en que se busca un equilibrio entre Precisión y Exhaustividad y la distribución desigual de clases es dada por muchos Negativos Actuales.

2.2. ÁRBOLES DE DECISIÓN

La idea de los árboles de decisión es representar en una estructura jerárquica todos los posibles casos del conjunto de entrenamiento para luego con la representación de alguna instancia determinar a qué clase pertenece.

El tipo de aprendizaje utilizado es Supervisado, el árbol aprende porque existe una clase que le dice constantemente que es correcto y que no.

El árbol posee dos tipos de nodos, nodos de decisión (que son los nodos internos) y los nodos hojas. Un nodo de decisión especifica alguna prueba sobre algún atributo particular. Un nodo hoja indica una clase.

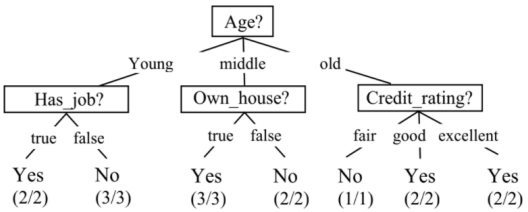

Fig 1. Ejemplo árbol de decisión [2].

En un árbol de decisión, los subconjuntos puros contienen ejemplos de entrenamiento de una sola clase, es decir, una hoja debería llegar con ejemplos de una sola clase.

Las funciones de impureza más populares para el aprendizaje de árboles de decisión son "ganancia de información" y "la tasa de ganancia de información", usados en C4.5 como dos alternativas.

2.2.1. GANANCIA DE INFORMACIÓN

La ganancia de información es una propiedad estadística que mide qué tan bien un atributo dado separa los ejemplos de entrenamiento de acuerdo con su clasificación objetivo [3].

Para tomar la elección de qué atributo será el próximo nodo del árbol de decisión, basándose en la ganancia de información que nos aportaría cada uno, debemos optar por escoger al que mayor resultado ofrezca. Para llegar a ello, se realiza una secuencia de cálculos previos, que consiste en:

1. Dado un conjunto de datos, se utiliza la función de entropía para calcular el valor de impureza del conjunto.

$$entropía\ (D)\ = -\sum_{j=1}^{c} Pr(cj)\ *\ log_2 Pr(cj) \qquad \textbf{(5)}$$

2. Luego se calcula la entropía de D respecto de cada atributo Ai, con el objetivo de conocer qué atributo reducirá más la impureza.

$$entropía_{Ai}(D) = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \times entropia(Dj). \quad \textbf{(6)}$$

3. Finalizando el procedimiento con el cálculo de la ganancia de la información de cada atributo Ai.

$$Ganancia(D, Ai) = entropía(D) - entropía(D). \quad \textbf{(7)}$$

Claramente, el criterio de ganancia de la información mide la reducción en la impureza o desorden. La entropía siempre será mejor cuando tienda a cero, por ende, la ganancia de información siempre será el valor más grande.

2.2.2. TASA DE GANANCIA DE INFORMACIÓN

La tasa de ganancia de información surge para corregir un defecto que presenta el método de ganancia de información para la selección del próximo nodo en árboles de decisión. Éste defecto se hace presente porque el criterio de ganancia tiende a favorecer o quedar sesgado por atributos con muchos valores posibles.

$$gainRatio(D, Ai) = \frac{ganancia(D,Ai)}{-\sum_{j=1}^{s} (\frac{|Dj|}{|D|} \times log_2 \frac{|Dj|}{|D|})} \quad \textbf{(8)}$$

La tasa de ganancia de información logra corregir este sesgo normalizando la ganancia usando la entropía de los datos con respecto a los valores del atributo. Es decir, que cuantos más posibles valores tenga el atributo Ai, el denominador será un número más grande y la ganancia que se obtendrá con la fórmula de ganancia (Ecuación 4) como la mejor, ahora se reducirá. ¿A quiénes beneficia esta expresión? A aquellos atributos que tengan menos cantidad de valores, compensando esa diferencia.

3. IMPLEMENTACIÓN

Con el objetivo de comparar los resultados obtenidos al utilizar Ganancia de información y Tasa de ganancia, se utilizarán un conjunto de datasets propios que buscarán reflejar a simple vista el problema que conlleva utilizar, para algunos casos, a la ganancia de información como criterio para la elección de nodos de decisión al generar un árbol. Para ello, se procederá a utilizar como métrica a la puntería (Ecuación 1) para medir el porcentaje de aciertos de clasificación de los conjuntos de prueba utilizados y en caso de ser necesario se optará por utilizar otras métricas: Precisión (Ecuación 2), Exhaustividad (Ecuación 3) y FScore (Ecuación 4), para poder evidenciar la precisión, confiabilidad de los modelos para clasificar nuevas instancias y el costo de clasificarlas mal.

Teniendo en cuenta como funciona el proceso de aprendizaje, se separará el conjunto de datos de los datasets que se utilizaran en dos: por una parte contaremos

con los datos de entrenamiento (utilizados para la generación del árbol) y por otra, con datos de prueba (utilizados para la validación del modelo).

4. PRUEBAS Y RESULTADOS

Para evidenciar el problema que conlleva utilizar la ganancia de información respecto de la tasa de información en la elección de los nodos durante la generación de árboles de decisión, se plantearon tres datasets con las siguientes características:

Tabla 2. Características de los datasets.

	Dataset 1	**Dataset 2**	**Dataset 3**	**Variabilidad**
N° registros	350	350	350	Baja
N° atributos	8	8	8	Media
N° de valores de clase	2	2	2	Alta

Con el objetivo de destacar como la ganancia de información muchas veces es sesgada por atributos con gran variabilidad de datos, optamos por tomar un atributo en común y atribuirle con la posibilidad de tomar una cantidad baja (2), media (8) y alta (20) de posibles valores en los datasets utilizados respectivamente.

Para comenzar las pruebas se optó por tomar un 70% (245/350) de los registros de cada dataset para utilizarse como datos de entrenamiento (utilizados para la generación del árbol de decisión) y el restante 30% (105/350) utilizarlo como datos de prueba para validar el modelo. A partir de la división de datos, se procedió a cargar los datasets al software desarrollado y así generar los árboles (tomando como criterio de decisión de nodos a la ganancia de información y tasa de ganancia por otro lado), para posteriormente poder validar los datos de prueba.

Tabla 3. Puntería al usar Ganancia de información.

	Dataset 1	**Dataset 2**	**Dataset 3**
Puntería	1	0.99	0.83

Tabla 4. Puntería al usar Tasa de ganancia.

	Dataset 1	**Dataset 2**	**Dataset 3**
Puntería	1	1	0.95

Una vez concluido el proceso de validación se tomaron esos datos para calcular la puntería (Ecuación 1) de los modelos, y visto que los resultados no fueron muy representativos (Tabla 3 y 4), se optó por tomar en cuenta otras métricas: Precisión, Exhaustividad y FScore, las cuales indicarán la precisión, confiabilidad de los modelos para clasificar nuevas instancias y el costo de clasificarlas mal. Para la obtención de estas nuevas métricas plantearon por cada dataset dos matrices de confusión (una

tomando a la Ganancia como criterio para elegir los nodos de decisión del árbol generado y otra matriz tomando la Tasa de ganancia).

Tabla 5. Matriz de confusión de dataset 1 tomando Ganancia de información.

	Negativo (NO)	**Positivo (SI)**
Negativo (NO)	71	0
Positivo (SI)	0	34

Tabla 6. Matriz de confusión de dataset 1 tomando Tasa de ganancia.

	Negativo (NO)	**Positivo (SI)**
Negativo (NO)	71	0
Positivo (SI)	0	34

Tabla 7. Matriz de confusión de dataset 2 tomando Ganancia de información.

	Negativo (NO)	**Positivo (SI)**
Negativo (NO)	69	1
Positivo (SI)	0	35

Tabla 8. Matriz de confusión de dataset 2 tomando Tasa de ganancia.

	Negativo (NO)	**Positivo (SI)**
Negativo (NO)	70	0
Positivo (SI)	0	35

Tabla 9. Matriz de confusión de dataset 3 tomando Ganancia de información.

	Negativo (NO)	**Positivo (SI)**
Negativo (NO)	74	8
Positivo (SI)	10	13

Tabla 10. Matriz de confusión de dataset 3 tomando Tasa de ganancia.

	Negativo (NO)	**Positivo (SI)**
Negativo (NO)	80	2
Positivo (SI)	0	23

Resumiendo los resultados de las métricas obtenidas para los tres datasets planteados utilizando a la ganancia de información como criterio para escoger los nodos en la generación de árboles de decisión:

Tabla 11. Resultados de utilizar Ganancia de información.

	Dataset 1	**Dataset 2**	**Dataset 3**
Puntería	1	0.99	0.83
Precisión	1	0.97	0.62
Exhaustividad	1	1	0.57
FScore	1	1	0.59

Resumiendo los resultados de las métricas obtenidas para los tres datasets planteados utilizando a la tasa de ganancia como criterio para escoger los nodos en la generación de árboles de decisión:

Tabla 12. Resultados de utilizar Tasa de ganancia.

	Dataset 1	**Dataset 2**	**Dataset 3**
Puntería	1	1	0.95
Precisión	1	1	0.92
Exhaustividad	1	1	1
FScore	1	1	0.96

5. ANÁLISIS

Para el Dataset 1, la poca variabilidad de sus atributos permite que los árboles generados usando como criterio la ganancia o la tasa de ganancia sean iguales, y además, la capacidad de predicción del modelo en base las pruebas realizadas es perfecta (punterías iguales a 1) por lo tanto al no tener ningún error de clasificación no sería necesario llevar a cabo otra métrica o análisis adicional. Sin embargo, vamos a analizar igualmente para ver a qué conclusiones llegamos a partir de los resultados obtenidos en el resto de métricas. Al observar la matriz de confusión podemos ver que no se obtuvieron falsos positivos ni falsos negativos (es decir, que cada clasificación realizada fue exitosa), de lo cual se derivan los siguientes resultados (Tabla 11 y 12):

- Precisión = 1 (en ambas pruebas) → no pago ningún costo por tener falsos positivos.
- Exhaustividad = 1 (en ambas pruebas) → no pago ningún costo por tener falsos negativos.
- FScore = 1 (en ambas pruebas) → como es la media armónica entre Precisión y Exhaustividad, se puede decir que los falsos positivos y falsos negativos no representan dificultades para medir la confiabilidad del modelo al momento de clasificar nuevas instancias. Y como se obtuvo el mismo resultado para ambas pruebas, es indiferente tomar al modelo con Tasa de ganancia o Ganancia de información.

En el Dataset 2, considerado de variabilidad media, al observar los resultados (Tabla 11 y 12) podemos destacar diferencias mínimas entre los valores obtenidos en cada métrica calculada. En lo que respecta a la puntería, se obtuvo un valor de 0.99 utilizando Ganancia de información, contra un 1 al usar Tasa de ganancia; una diferencia mínima que da a lugar a que haya solo un 1% de posibilidades de que el modelo clasifique mal a una nueva instancia al utilizar Ganancia de información.

El cálculo de las métricas Precisión, Exhaustividad y FScore para la Ganancia ya es un valor distinto a 1, pero aun no es de utilidad llevar a cabo sus cálculos, ya que no

se observa una cantidad muy sesgada en la distribución de valores posible para la Clase en el conjunto de prueba. Sin embargo, se detallará a continuación los detalles de haberlos calculado con el objetivo de hacer notar ésta aclaración. Al observar la matriz de confusión obtenida, vemos que para la Ganancia de información no se obtuvieron falsos positivos ni falsos negativos (es decir, que cada clasificación realizada fue exitosa). Al contrario de la matriz obtenida para la Tasa de ganancia en la cual se encontró 1 falso positivo y ningún falso negativo. Al ser una cantidad tan baja de falsos positivos, obtenemos que:

- Precisión = 1 (Tasa de ganancia) → no pago ningún costo por tener falsos positivos.
- Precisión = 0.97 (Ganancia de inf.) → al no ser un valor tan lejano a 1, la presencia de falsos positivos no representa un costo significativo, por lo cual es casi indiferente tomar al modelo con tasa o ganancia.
- Exhaustividad = 1 (en ambos casos) → no pago ningún costo por tener falsos negativos.
- FScore = 1 (en ambos casos) → como es la media armónica entre Precisión y Exhaustividad, entonces los falsos positivos y los falsos negativos no representan dificultades para medir la confiabilidad del modelo al momento de clasificar nuevas instancias. Y como se obtuvo el mismo resultado para ambas pruebas, es indiferente nuevamente tomar al modelo con Tasa de ganancia o Ganancia de información.

En el dataset 3, cuya variabilidad de atributos es alta, los valores de las métricas obtenidas fueron variando respecto a los resultados obtenidos con la Ganancia de información y Tasa de ganancia. Obteniendo una notoria diferencia entre las punterías (0.83 en Ganancia de inf. contra 0.95 en Tasa de ganancia), lo cual nos indica que el porcentaje de aciertos es menor y que igualmente existe una probabilidad del 5% de que el modelo haga malas clasificaciones al utilizar la tasa de ganancia. Cabe mencionar que se puede observar la existencia de un importante sesgo entre la cantidad uno de los valores (NO) posibles de la clase respecto al otro (SI). Para el conjunto de prueba, se clasificaron 84 veces con NO y respecto al valor SI 21 veces, lo cual nos indica que la Puntería no es una métrica tan representativa para elegir un modelo, es por ello que se decide usar la Precisión, Exhaustividad y FScore para tomar la decisión (escogiendo aquella que su valor obtenido sea más alto). Estos valores son provenientes de la Matriz de Confusión [10] generada a partir de los valores de la clase. Al observar la matriz de confusión de Tasa de ganancia, observamos que se presentan 2 falsos positivos y ningún falso negativo, al contrario de la Ganancia de información que presenta 8 falsos positivos y 10 falsos negativos; con lo cual obtenemos:

- Precisión = 0.62 (Ganancia de información) → se paga un costo bastante alto por tener falsos positivos.

- Precisión = 0.92 (Tasa de ganancia) → se paga un costo leve por tener falsos positivos.
- Exhaustividad = 0.57 (Ganancia de información) → se paga un alto costo por falsos negativos.
- Exhaustividad = 1 (Tasa de ganancia) → no se paga ningún costo por tener falsos negativos.
- FScore = 0.59 para Ganancia de inf. y 0.96 para Tasa de ganancia → El equilibrio entre Precisión y Exhaustividad es mucho mayor al utilizar Tasa de ganancia respecto de usar Ganancia de información.

6. CONCLUSIÓN

En este estudio nos enfocamos en los Árboles de Decisiones, y a través de la implementación de su algoritmo usando un software propio pudimos demostrar (de forma práctica) que:

- Para datasets con atributos que poseen una poca cantidad de valores posibles, el uso de Ganancia de Información y Tasa de Ganancia para la selección de nodos es indistinto, ambos generan árboles con una confiabilidad muy alta, (la mayoría de las veces generan el mismo árbol).
- Para datasets complejos que poseen atributos con gran variabilidad de valores posibles, la Tasa de Ganancia para la selección de nodos genera árboles de mayor confiabilidad (ya que posee la capacidad de normalizar los datos de estos atributos y evitar valores poco representativos de la entropía), es decir brinda clasificaciones más confiables ante la incorporación de nuevas instancias.

REFERENCIAS

1. Quinlan, J. C4. 5: programs for machine learning. 1993: Morgan Kaufmann Publishers.
2. Bing Liu. Web Data Mining. Exploring Hyperlinks, Contents, and Usage Data. (2008).
3. Árbol de decisión en Machine Learning (Parte 1). (2019, December 14). sitiobigdata.com. Retrieved November 27 (2021)
4. Accuracy, Precision, Recall or F1? | by Koo Ping Shung. (n.d.). Towards Data Science. Retrieved December 4 (2021)