Algoritmo ID3

Miguel Rodríguez Asensio

1 de julio de 2022

NOTA. Este material es optativo y está desarrollado para que el alumno **que así lo desee** pueda profundizar en la materia explicada en el vídeo de este apartado.

El algoritmo ID3 (**I**terative **D**ichotomizer **3**) fue ideado por Quinlan en 1986, que después evolucionó a C4.5 (Quinlan 1993) y C5.0 (Quinlan 1997). El algoritmo ID3 de forma resumida, consiste en seleccionar un atributo como nodo raíz del árbol y entonces crear una rama con cada uno de los valores de ese atributo. El proceso es iterativo, de modo que con cada rama resultante se realiza el mismo proceso, es decir, se selecciona otro atributo y se genera una nueva rama por cada posible valor del atributo. El proceso continua hasta que los ejemplos se clasifiquen a través de uno de los caminos del árbol.

Para desarrollar este método nos vamos a basar en el siguiente conjunto de datos:

Vista	Temperatura	Humedad	Viento	Jugar
Soleado	Alta	Alta	No	No
Soleado	Alta	Alta	Si	No
Nublado	Alta	Alta	No	Si
Lluvioso	Media	Alta	No	Si
Lluvioso	Baja	Normal	No	Si
Lluvioso	Baja	Normal	Si	No
Nublado	Baja	Normal	Si	Si
Soleado	Media	Alta	No	No
Soleado	Baja	Normal	No	Si
Lluvioso	Media	Normal	No	Si
Soleado	Media	Normal	Si	Si
Nublado	Media	Alta	Si	Si
Nublado	Alta	Normal	No	Si
Lluvioso	Media	Alta	Si	No

A efectos de aclarar la terminología utilizada en lo que sigue, vamos a llamar atributos a las variables Vista, Temperatura, Humedad y Viento, y clase a la variable Jugar (que tendrá un total de dos clases o modalidades).

Definiciones:

Definimos la **entropía** como:

$$I(A) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

El criterio de partición utilizado por este algoritmo ID3 está basado en **la ganancia de la información** que se obtiene en base a la entropía, y la idea es medir cuanta información se gana al particionar por un determinado atributo (o variable independiente). Esta ganancia de la información se define como la diferencia de entropía del nodo actual y la suma ponderada de las entropías correspondientes a bifurcar por ese atributo.

Entonces se define la ganancia de información al usar el tributo A_i como:

$$G(A_i) = I - I(A_i) \tag{1}$$

siendo I la información (entropía) antes de utilizar el atributo e $I(A_i)$ la información después de usarlo. Los valores anteriores se obtienen de la siguiente manera:

$$I = -\sum_{c=1}^{nc} \frac{n_c}{n} log_2 \left(\frac{n_c}{n}\right) \tag{2}$$

$$I(A_i) = \sum_{j=1}^{nv(A_i)} \frac{n_{ij}}{n} I_{ij} \; con \; I_{ij} = -\sum_{k=1}^{nc} \frac{n_{ijk}}{n_{ij}} log_2 \left(\frac{n_{ijk}}{n_{ij}}\right) \tag{3}$$

La notación empleada en las fórmulas anteriores es:

- nc, es el número de clases (2 en nuestro ejemplo).
- n_c es el número de ejemplares de la clase c. Es decir 5 para el valor NO y 9 para el valor SI de la variable jugar y en el momento de arranque del algoritmo.
- El número del valores del atributo A_i, se representa por $nv(A_i)$. nv(Vista)=3, nv(Temperatura)=3, nv(Humedad)=2, nv(Viento)=2 al comienzo del algoritmo
- El número de ejemplos (observaciones) del valor j dentro del atributo A_i, se representa por n_{ij}. Por ejemplo para el tributo Vista, $n_{11} = 5$ ($soleado$); $n_{12} = 4$ ($Nublado$); $n_{13} = 5$ ($lluvioso$), en el arranque del algoritmo.
- Por n_{ijk} representamos el número de observaciones o ejemplos con valor j dentro de A_i que pertenecen a la clase k. Así $n_{111} = 3$ para la variable vista con valor soleado y la clase jugar=NO

Para ver con mayor claridad y detalle este procedimiento, procedemos a continuación a desarrollar este algoritmo para el ejemplo planteado al principio, es decir vamos a decidir si jugar al tenis o no en base a los atributos relativos al tiempo que hace. Vamos a seleccionar en un primer paso el atributo que maximice la ganancia, por lo que en un principio calculamos la entropía para la variable Jugar

$$I = -\frac{5}{14} lg_2 \left(\frac{5}{14}\right) - \frac{9}{14} lg_2 \left(\frac{9}{14}\right) = 0{,}9403$$

En este caso:

- nc = número de clases = 2 (Jugar SI ó NO)

- n_c= número de ejemplos en cada valor de la clase (9 en SI y 5 en NO)
- n = número de ejemplos = 14

Ahora utilizamos la ecuación 1, para calcular la ganancia de información que tenemos con cada uno de los atributos: vista, temperatura, humedad y viento.

Comenzamos con al atributo Vista.

$$G(Vista) = I - I(Vista)$$

Calculamos la entropía para cada uno de los posibles valores del atributo vista. Para mejorar la interpretación de las fórmulas utilizadas hacemos la siguiente clasificación.

		Jugar		Total
		SI	NO	
Vista	Soleado	2	3	5
	Nublado	4	0	4
	Lluvioso	3	2	5

Entonces, para el valor de soleado en el atributo vista, se tiene

$$I_{vista,soleado} = -\left(\frac{2}{5}lg_2\left(\frac{2}{5}\right) + \frac{3}{5}lg_2\left(\frac{3}{5}\right)\right) = 0{,}9709$$

$$I_{vista,Nublado} = -\left(\frac{4}{4}lg_2\left(\frac{4}{4}\right) + \frac{0}{4}lg_2\left(\frac{0}{4}\right)\right) = 0$$

$$I_{vista,Lluvisos} = -\left(\frac{3}{5}lg_2\left(\frac{3}{5}\right) + \frac{3}{5}lg_2\left(\frac{2}{5}\right)\right) = 0{,}9709$$

Por lo tanto:

$$I(Vista) = \frac{5}{14}0{,}9709 + \frac{4}{14}0 + \frac{5}{14}0{,}9709 = 0{,}6935$$

En consecuencia

$$G(Vista) = 0{,}9403 - 0{,}6935 = 0{,}2468$$

Para estos cálculos se ha tenido en cuenta que:

- nv_{Vista}=número de valores de la variable vista = 3
- n_{ij} = número de ejemplos con el valor j (Soleado 5, Nublado 4, Lluvioso 5)
- n_{ijk} = número de ejemplos con valor j en A_i y que pertenecen a la clase k (Soleado: SI 2, NO 3; Nublado: SI 4, NO 0; Lluvioso: SI 3, NO 2)

De forma similar se obtendrá la ganancia de información para el resto de atributos, obteniendo los siguiente:

- G(Temperatura)= 0.0292
- G(Humedad)=0.1519
- G(Viento)=0.0481

Por tanto, en el primer nodo del árbol se decide que el mejor atributo para el nodo es Vista, debido a que presenta un mayor valor de G (ganancia de información). A continuación, se generan nodos para cada valor del atributo y, en el caso de Vista=Nublado se llega a un nodo hoja, ya que todos los ejemplos de entrenamiento que llegan a dicho nodo son de la misma clase SI. Sin embargo, para los otros dos casos se repite el proceso de elección con el resto de atributos y con los ejemplos de entrenamiento que se clasifican a través de ese nodo.

Como resumen de todo lo visto hasta ahora, se tiene que el primer nodo que entraba es «Vista». Además en el caso de *Vista=nublado* ya se llega a un nodo hoja. Ahora queda estudiar el resto de los valores que puede tomar el atributo Vista.

a) **caso «Vista=Soleado»**.

En este caso la tabla que queda es la siguiente

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
1	Soleado	Alta	Alta	NO	NO
2	Soleado	Alta	Alta	SI	NO
8	Soleado	Media	Alta	NO	NO
9	Soleado	Baja	Normal	NO	SI
11	Soleado	Media	Normal	SI	SI

En primer lugar calculamos I:

$$I = -(3/5) * log_2(3/5) - (2/5) * log_2(2/5) = 0{,}9709506$$

Ahora tendremos que calcular G(Temperatura), G(Humedad), G(viento)

G(Temperatura)=I-I(Temperatura)

$$I_{temperatura,Alta} = -(2/2 * log_2(2/2) + 0 * lg_2(0)) = 0$$

$$I_{temperatura,Media} = -(1/2 * log_2(1/2) + 1/2 * log_2(1/2)) = 1$$

$$I_{temperatura,Baja} = (1 * lg_2(1) + 0 * log_2(0)) = 0$$

Por lo tanto[1]

$$I(Temperatura) = (2/5) * 0 + (2/5) * 1 + (1/5) * 0) = 2/5 = 0{,}4$$

Luego $G(Temperatura) = 0{,}9709506 - 0{,}4 = 0{,}5709506$
G(Humedad)=I-I(Humedad)

$$I_{Humedad,Alta} = 0$$

$$I_{Humedad,Normal} = 0$$

De aquí se obtiene que:

$$I(Humedad) = 0$$

Luego $G(Humedad) = 0{,}9709506$

Por lo tanto aquí entraría humedad, que para «Alta» se elegirá NO y para «Normal», se elegirá SI. Por aquí ya se ha llegado a un nodo de tipo hoja, cuando se hace la ramificación.

b) **caso «Vista=Lluvioso»**.

La tabla queda como sigue

Ejemplo	Vista	Temperatura	Humedad	Viento	Jugar
4	Lluvioso	Media	Alta	NO	SI
5	Lluvioso	Baja	Normal	NO	SI
6	Lluvioso	Baja	Normal	SI	NO
10	Lluvioso	Media	Normal	NO	SI
14	Lluvioso	Media	Alta	SI	NO

[1]Observar que en los casos en los que se tiene I=0, no sería necesario hacer operación alguna, ya que en estos casos los datos pertenecen a una sola clase.

En este caso se puede ver que si se elige «viento» se va a llegar a una hoja después de la ramificación, ya que en todos los casos en que esta variable toma el valor «NO» están asociados con SI de la variable Jugar, y todos los casos con «NO» están asociados con SI de la variable Jugar. No obstante, vamos a calcular para este atributo el valor de la función G para confirmar esto.

Primero calculamos I:

$$I = -(3/5) * log_2(3/5) - (2/5) * log_2(2/5) = 0{,}9709506$$

Ahora tendremos que calcular G(Temperatura), G(viento). Calculemos G(viento) para ver que es el máximo, es decir coincide con I.

G(viento)=I-I(viento)

$$I_{viento,NO} = -((3/3) * log_2(3/3) + 0 * log_2(0)) = 0$$

$$I_{viento,SI} = 0$$

Por lo tanto G(viento)=I-0=I, y en consecuencia el atributo que se añade es «Viento», que como ya se ha dicho antes, es también un nodo hoja después de la ramificación, y por lo tanto el árbol estaría completado pues todos los nodos finales son hoja, que es lo que se pretende con este algoritmo.

El árbol final que nos queda sería el siguiente.