Instituto Politécnico Nacional
Escuela Superior de Cómputo
Secretaría Académica
Departamento de Ingeniería en Sistemas Computacionales

Minería de datos (*Data Mining*)
Medidas de particionamiento en árboles de decisión

1

Profesora: *Dra. Fabiola Ocampo Botello*

Bhumika, Aditya, Akshay, Arpit & Naresh (2017) diversas medidas para la selección de atributos para dividir las tuplas en un árbol.

La medida de selección de atributos determina cómo dividir las tuplas en un nodo dado y, por lo tanto, también se conocen como reglas de división.

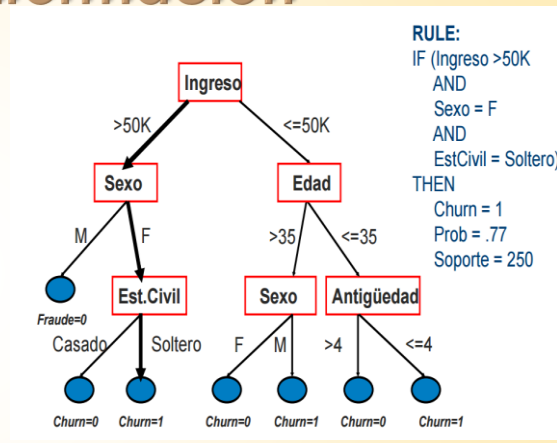
El nodo de árbol para la partición está etiquetado con el criterio de división, las ramas se generan para cada resultado del criterio y las tuplas se dividen en consecuencia.

Las medidas de selección de atributos más populares son: entropía (ganancia de información), relación de ganancia e índice de Gini.

2

Ganancia de Información

Sancho Capparini, Fernando (2009) indica que el **árbol ID3** usa el concepto de **Ganancia de Información** para seleccionar el atributo más útil en cada paso. Utiliza un método voraz para decidir la pregunta que mayor ganancia proporcione en cada paso, esto es, aquella que permite separar mejor los ejemplos respecto a la clasificación final.



La estrategia básica del ID3 es elegir los atributos de particionamiento con la mayor información.

Entropía (*Entropy*)

La entropía es una medida de incertidumbre asociada con una variable aleatoria. La entropía aumenta con el aumento de la incertidumbre o aleatoriedad y disminuye con una disminución de la incertidumbre o aleatoriedad. El valor de la entropía varía de 0 a 1.

$$Entropía(D) = E(D) = \sum_{i=1}^{c} -pi\ Log_2\ (pi)$$

donde pi es la probabilidad distinta de cero de que una tupla arbitraria en D pertenezca a la clase C y se estima mediante |Ci,D|/|D|. Se utiliza una función de registro de la base 2 porque, como se indicó anteriormente, la entropía está codificada en los bits 0 y 1. (Bhumika et al., 2017)

El concepto usado para cuantificar la información se llama **entropía**. *La entropía es usada para medir la cantidad de incertidumbre en un conjunto de datos.*

Sancho Capparini, Fernando (2009) presenta dos ejemplos para comprender la incertidumbre:

1. En una muestra totalmente homogénea, en la que todos los elementos se clasifican por igual tiene una incertidumbre mínima, esto es, no se tienen dudas de cuál es la clasificación de cualquiera de sus elementos. En este caso la incertidumbre (entropía) es cero.

5

2. En una muestra igualmente distribuida en el que se tienen el mismo número de casos en cada posible clasificación tiene una incertidumbre máxima. En este caso, la incertidumbre (entropía) es 1.

Ganancia de Información (Information Gain)

ID3 utiliza la ganancia de información como su medida de selección de atributos.

La ganancia de información es la diferencia entre el requisito de ganancia de información original (es decir, basado solo en la proporción de clases) y el nuevo requisito (es decir, obtenido después de la división de A). (Bhumika, et al., 2017).

$$Gain(D,A)=Entropy(D)-\sum_{j=1}^{v}\frac{|Dj|}{|D|}Entropy(D_j)$$

6

Dónde,
D: una partición de datos dada
A: atributo
V: Supongamos que dividimos las tuplas en D en algún atributo A que tiene v valores distintos

D se divide en v partición o subconjuntos, {D1, D2, .. Dj} donde Dj contiene esas tuplas en D que tienen el resultado aj de A.
Se elige el atributo que tiene la mayor ganancia de información.

Cuando todos los elementos pertenecen a la misma clase, la entropía es cero. La intención de un árbol es realizar particiones del conjunto de datos con la finalidad de que todos los elementos del subconjunto final pertenezcan a la misma clase (pureza).

La incertidumbre es máxima cuando los dos eventos tienen la misma probabilidad de ocurrencia.

7

ID3 elige el atributo de división con la **mayor ganancia de información**, donde la ganancia se define como la diferencia entre cuánta información se necesita para hacer una clasificación correcta antes de la división y cuánta información se necesita después de la división (Dunham, 2002).

p(0.3, 0.7)

p(0.5, 0.5)

p(0.2, 0.8)

p(0, 1.0)

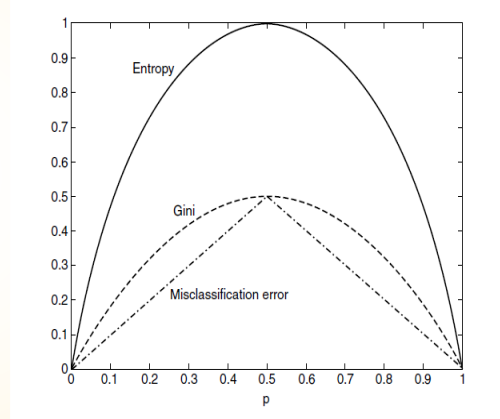


Figure 4.13. Comparison among the impurity measures for binary classification problems.

Tan et al., (2005) establecen que para determinar la bondad de una condición de prueba de atributo, necesitamos comparar el grado de impureza del nodo primario (antes de dividir) con el grado ponderado de impureza de los nodos secundarios (después de dividir).

Cuanto mayor sea su diferencia, mejor será la condición de la prueba. Esta diferencia, Δ, también denominada ganancia de pureza de una condición de prueba de atributo, se puede definir de la siguiente manera:

$$\Delta = I(\text{parent}) - I(\text{children})$$

8

El algoritmo de aprendizaje del árbol de decisión selecciona la condición de prueba de atributo que muestra **la máxima ganancia**.

Los pasos a seguir son los siguientes (Sancho Capparini, Fernando (2009)):

Paso	Descripción
1	Se calcula la entropía total. $$E(s) = \sum_{i=1}^{c} -pi\ Log_2\ (pi)$$
2	Se divide el conjunto de datos en términos de los diversos atributos.
3	Se calcula la entropía de cada rama y se suman proporcionalmente las ramas para calcular la entropía del total $$E(T,X) = \sum_{c \in x} p(c)E(Sc)$$
4	Se resta este resultado de la entropía original, se obtiene como resultado la Ganancia de Información (descenso de entropía) usando este atributo. $$Gain(T,X) = E(T) - E(T,X)$$
5	El atributo con mayor Ganancia se elige como nodo de decisión.

9

Ejemplo de la aplicación de la ganancia de información y entropía (ejemplo tomado de Sancho Capparini, Fernando (2009)). Aplicado al conjunto de datos del juego de Golf.

JuegaGolf	Panorama	Temperatura	Humedad	Viento
No	Lluvioso	Caliente	Alta	FALSO
No	Lluvioso	Caliente	Alta	VERDADERO
Si	Nublado	Caliente	Alta	FALSO
Si	Soleado	Templado	Alta	FALSO
Si	Soleado	Frío	Normal	FALSO
No	Soleado	Frío	Normal	VERDADERO
Si	Nublado	Frío	Normal	VERDADERO
No	Lluvioso	Templado	Alta	FALSO
Si	Lluvioso	Frío	Normal	FALSO
Si	Soleado	Templado	Normal	FALSO
Si	Lluvioso	Templado	Normal	VERDADERO
Si	Nublado	Templado	Alta	VERDADERO
Si	Nublado	Caliente	Normal	FALSO
No	Soleado	Templado	Alta	VERDADERO

Jugar Golf	
SI	NO
9	5

10

PASO 1. Cálculo de la entropía total

$$E(s) = \sum_{i=1}^{c} -pi\ Log_2\ (pi)$$

Jugar Golf	
SI	NO
9	5

11

E(Jugar Golf) = E(No, Si) = E(5,9)

= (-5/14 Log_2 (5/14)) + (-9/14 Log_2 (9/14))

= (-0.36 Log_2 (0.36)) + (-0.64 Log_2 (0.64))

= 0.53 + 0.40 = **0.94 --> Entropía total**

PASO 2. Dividir el conjunto de datos en los diversos atributos.

Atributo Objetivo:	Jugar Golf
Atributo	**Dominio**
Panorama	Lluvioso Nublado Soleado
Temperatura	Caliente Frío Templado
Humedad	Normal Alta
Viento	Falso Verdadero

12

PASO 3. Se calcula la entropía en cada rama y se suman proporcionalmente para calcular la entropía total:

$$E(T,X) = \sum_{c \in x} p(c)E(C)$$

				Count
Panorama	Lluvioso	JuegaGolf	No	3
			Si	2
	Nublado	JuegaGolf	Si	4
	Soleado	JuegaGolf	No	2
			Si	3

E(Soleado) = E(No, Si) = E(2,3) = (-2/14 Log_2 (2/14)) + (-3/14 Log_2 (3/14))
= (-0.14 Log_2 (0.14)) + (-0.21 Log_2 (0.21))
= 0.40 + 0.47 = **0.97 --> Entropía Soleado**

Para panorama:

E(Jugar Golf, Panorama) = P(Lluvioso) * E(3,2) + P(Nublado) * E(4,0) + P(Soleado) * E(2,3)
-Lluvioso: P(5/14) = 0.36 E(3,2) = 0.44 + 0.53 = 0.971
-Nublado: P(4/14) = 0.29 E(4,0) = 0
-Soleado: P(5/14) = 0.36 E(2,3) = 0.971
E(Jugar Golf, Panorama) = 0.36 * 0.971 + 0.29 * 0 + 0.36 * 0.971 = 0.35 + 0 + 0.35 = 0.70
GAIN = 0.94 - 0.70 = 0.247

13

PASO 4. Ganancia de información

$$Gain(T,X) = E(T) - E(T,X)$$

Para temperatura:

E(Jugar Golf, Temperatura) = P(Caliente) * E(2,2) + P(Frío) * E(1,3) + P(Templado) * E(2,4)

-Caliente: P(4/14) = 0.29 E(2,2) = 0.5 + 0.5 = 1
-Frío: P(4/14) = 0.29 E(1,3) = 0.5 + 0.31 = 0.81
-Templado: P(6/14) = 0.43 E(2,4) = 0.53 + 0.39 = 0.92

E(Jugar Golf, Temperatura) = 0.29 * 1 + 0.29 * 0.81 + 0.43 * 0.92 = 0.29 + 0.2349 + 0.39 = 0.91

GAIN = 0.94 - 0.91 = 0.03

14

				Count
Temp	Caliente	JuegaGolf	No	2
			Si	2
	Frio	JuegaGolf	No	1
			Si	3
	Templado	JuegaGolf	No	2
			Si	4

PASO 4. Ganancia de información

Para humedad:

E(Jugar Golf, Humedad) = P(Normal) * E(1,6) + P(Alta) * E(4,3)

-Normal: P(7/14) = 0.50 E(1,6) = 0.40 + 0.19 = 0.59
-Alta: P(7/14) = 0.50 E(4,3) = 0.46 + 0.52 = 0.98

E(Jugar Golf, Humedad) = 0.50 * 0.59 + 0.50 * 0.98 = 0.295 + 0.49 = 0.785

GAIN = 0.94 - 0.785 = 0.155

PASO 4. Ganancia de información

15

				Count
Humedad	Alta	JuegaGolf	No	4
			Si	3
	Normal	JuegaGolf	No	1
			Si	6

Para viento:

E(Jugar Golf, viento) = P(Falso) * E(6,2) + P(Verdadero) * E(3,3)

-Falso: P(8/14) = 0.57 E(6,2) = 0.31 + 0.50 = 0.81
-Verdadero: P(6/14) = 0.43 E(3,3) = 0.50 + 0.50 = 1

E(Jugar Golf, Viento) = 0.57 * 0.81 + 0.43 * 1 = 0.46 + 0.43 = 0.89

GAIN = 0.94 - 0.89 = 0.05

PASO 4. Ganancia de información

16

				Count
Viento	Falso	JuegaGolf	No	2
			Si	6
	Verdadero	JuegaGolf	No	3
			Si	3

PASO 5. Elegir el atributo con mayor Ganancia de información

Variable	Ganancia	(Si, No)
Panorama	0.247	Lluvioso = (3,2). Nublado = (4,0). Soleado = (2,3)
Temperatura	0.03	Caliente = (2,2). Frío = (1,3). Templado = (2,4)
Humedad	0.155	Normal = (1,6). Alta = (4,3)
Viento	0.05	Falso = (6,2). Verdadero = (3,3)

17

Panorama es la variable que brinda la mayor ganancia de información, por tal, será la primera en ser elegida.

Gain Ratio

Bhumika, et al (2017) establece que la **medida de ganancia de información** está sesgada hacia pruebas con muchos resultados. Es decir, prefiere seleccionar atributos que tengan una gran cantidad de valores. Como cada partición es pura, la ganancia de información por partición es máxima. Pero tal partición no puede usarse para la clasificación.

C4.5 (un sucesor de ID3) utiliza esta medida de selección de atributo denominada Gain Ratio, que es una extensión de la ganancia de información (Bhumika, et al., 2017).

18

$$SplitInfo_A = -\sum_{j=1}^{v} \frac{|Dj|}{|D|} log_2\left(\frac{|Dj|}{|D|}\right)$$

La relación de ganancia (Gain Ratio) se define entonces como:

$$Gain\ Ratio(A) = \frac{Gain\ (A)}{SplitInfoA\ (D)}$$

Gini Index

El índice de Gini es una medida de selección de atributos utilizada por el algoritmo del árbol de decisiones **CART**. El índice de Gini mide la impureza D, una partición de datos o un conjunto de tuplas de entrenamiento como (Bhumika et al., 2017):

$$Gini(D)=1-\sum_{i=1}^{m} p_i^2$$

19

Donde pi es la probabilidad de que una tupla en D pertenezca a la clase Ci y se estima mediante |Ci, D|/|D|. La suma se calcula sobre m clases. El atributo que reduce la impureza al nivel máximo (o tiene el índice mínimo de Gini) se selecciona como el atributo de división (Bhumika et al., 2017).

Ejemplo propuesto por: Bhumika, et al (2017).

Ejemplo de la base de datos de una tienda electrónica para ver si una persona compra una computadora portátil o no.

ID	age	salary	graduate	credit_rating	class: buys_laptop
1	youth	high	no	average	no
2	youth	high	no	excellent	no
3	middle_aged	high	no	average	yes
4	senior	medium	no	average	yes
5	senior	low	yes	average	yes
6	senior	low	yes	excellent	no
7	middle_aged	low	yes	excellent	yes
8	youth	medium	no	average	no
9	youth	low	yes	average	yes
10	senior	medium	yes	average	yes
11	youth	medium	yes	excellent	yes
12	middle_aged	medium	no	excellent	yes
13	middle_aged	high	yes	average	yes
14	senior	medium	no	excellent	no

Figure 1. Class-labeled labeled training tuples from the Electronics Store database.

El atributo con etiqueta de clase compra _laptop, tiene dos valores distintos (sí, no). Por lo tanto, hay dos clases distintas y el valor de m es igual a 2.

Clase P: buys_laptop = "yes"
Clase N: buys_laptop = ”no”

Como hay 9 sí y 5 no en el atributo buys_laptop, por lo tanto, 9 tuplas pertenecen a la clase P y 5 tuplas pertenecen a la clase N.

Entropy is calculated as:

$$Entropy(D) = -\frac{9}{14}log_2\left(\frac{9}{14}\right) - \frac{5}{14}log_2\left(\frac{5}{14}\right) = 0.940$$

Now the information gain is calculated as:

$$Gain(age,D) = Entropy(D) - \sum_{v \in \{youth, middle-aged, Senior\}} \frac{|Sv|}{|S|} Entropy(S_v)$$

$$= Entropy(D) - \frac{5}{14}Entropy(S_{youth}) - \frac{4}{14}Entropy(S_{middle-aged}) - \frac{5}{14}Entropy(S_{senior})$$

$$Gain(salary,D) = 0.029$$

$$Gain(graduate,D) = 0.151$$

$$Gain(credit\_raiting,D) = 0.048$$

Imagen tomada del artículo:

Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, Naresh Dhami. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications (0975–8887). Volume 163 – No 8, April 2017.

Tan et al. (2005) establece que hay medidas que pueden usarse para determinar la bondad de una condición de prueba de un atributo. Estas medidas intentan dar preferencia a las condiciones de prueba de atributos que dividen las instancias de entrenamiento en subconjuntos puros en los nodos secundarios.

La impureza de un nodo mide qué tan diferentes son las etiquetas de clase para las instancias de datos que pertenecen a un nodo común.

Las medidas para evaluar la impureza de un nodo:

$$\begin{aligned} \text{Entropy} &= -\sum_{i=0}^{c-1} p_i(t)\log_2 p_i(t), \\ \text{Gini index} &= 1 - \sum_{i=0}^{c-1} p_i(t)^2, \\ \text{Classification error} &= 1 - \max_i[p_i(t)], \end{aligned}$$

Imagen tomada de Tan et al., (2005)

Ejemplos de cálculo de impureza:

Node N_1	Count
Class=0	0
Class=1	6

$\text{Gini} = 1 - (0/6)^2 - (6/6)^2 = 0$
$\text{Entropy} = -(0/6)\log_2(0/6) - (6/6)\log_2(6/6) = 0$
$\text{Error} = 1 - \max[0/6, 6/6] = 0$

Node N_2	Count
Class=0	1
Class=1	5

$\text{Gini} = 1 - (1/6)^2 - (5/6)^2 = 0.278$
$\text{Entropy} = -(1/6)\log_2(1/6) - (5/6)\log_2(5/6) = 0.650$
$\text{Error} = 1 - \max[1/6, 5/6] = 0.167$

Node N_3	Count
Class=0	3
Class=1	3

$\text{Gini} = 1 - (3/6)^2 - (3/6)^2 = 0.5$
$\text{Entropy} = -(3/6)\log_2(3/6) - (3/6)\log_2(3/6) = 1$
$\text{Error} = 1 - \max[3/6, 3/6] = 0.5$

Ejemplo tomado de Tan et al., (2005)

Rokach & Maimon (2015) establecen los siguientes criterios de detención del crecimiento:

- Todas las instancias en el conjunto de entrenamiento pertenecen a un solo valor de *y* (clase).
- Se ha alcanzado la profundidad máxima del árbol.
- El número de casos en el nodo terminal es menor que el número mínimo de casos para los nodos principales (padres).

Los mismos autores señalan que evaluar el desempeño de un árbol de clasificación es una tarea fundamental en el aprendizaje automático.

Algunos de los criterios que presentan son:
- La matriz de confusión, la cual presenta la cantidad de elementos que han sido clasificados correcta e incorrectamente.
- El coeficiente de correlación.

Poda de árboles

Rokach & Maimon (2015) establecen que el empleo de criterios estrictos de detención tiende a crear árboles de decisión pequeños y mal balanceados. Por otro lado, el uso de otros criterios de detención tiende a generar grandes árboles de decisión que están sobreajustados para el conjunto de entrenamiento.

Breiman *et al.* (1984, citado en Rokach & Maimon, 2015) desarrolló una metodología de poda basada en un criterio de detención que permite que el árbol de decisión sobreajuste el conjunto de entrenamiento, en donde el árbol sobreajustado se corta en un árbol más pequeño eliminando las subramas que no contribuyen a la precisión de la generalización.

Referencias bibliográficas

Bhumika Gupta, Aditya Rawat, Akshay Jain, Arpit Arora, Naresh Dhami. (2017). Analysis of Various Decision Tree Algorithms for Classification in Data Mining. International Journal of Computer Applications (0975–8887). Volume 163 – No 8, April 2017.

Dunham, M. H. (2002). *Data mining: introductory and advanced topics*. Prentice Hall.

Rokach, L. & Maimon, O. (2015). Data Mining with decision trees. Theory and Applications. Second Edition. World Scientific Publishing Co. Pte. Ltd.

Sancho Capparini, Fernando (2009). Aprendizaje inductivo. Arboles de decisión. Portal Web. Disponible en: http://www.cs.us.es/~fsancho/?e=104

Tan Pang-Ning, Steinbach Michael, Kumar Vipin. (2005). Introduction to data mining. First Edition. Pearson New International Edition.