Universidad de Santiago de Chile
Facultad de Ingeniería
Depto. de Ingeniería Informática

Minería de Datos
Capítulo VII
"Árboles de Decisión"

Profesor: Dr. Max Chacón.

Objetivos

- **Comprender la generación de un árbol de decisión.**
- **Cuantificar la ganancia de información para un atributo en un conjunto de datos.**
- **Comprender los algoritmos de generación de los árboles de decisión.**
- **Establecer los mecanismos de poda de los árboles de decisión.**
- **Comprender los mecanismos de equivalencia de reglas y la generalización de reglas simples.**

7.1.Definiciones.

Los árboles de decisión fueron presentados por J. R. Quinlan en 1983, se verá la versión C4.5. Actualmente existen variaciones de estos algoritmos.

La idea original se basa en los trabajos de Hoveland y Hunt de 1950, y Hunt, Marin y Stone 1966, los cuales están basados en modelos psicológicos de como las personas aprenden conceptos simples.

La idea básica es lo que se denomina *sistemas de aprendizaje conceptual*, los cuales intentan distinguir características de un conjunto de entrenamiento, que es en esencia una aplicación del método de *divide y conquista*.

Considerando una base de datos o sistema de información procedimental $SI=<U, Q, V, f>$ como fue definida:

- $S \subseteq U$ universo cerrado: un conjunto finito, no vacío, de n *objetos* $\{x_1, x_2, \ldots, x_n\}$
- Q: un conjunto finito, no vacío, de p *atributos* $\{q_1, q_2, \ldots, q_p\}$
- $V=\bigcup_{q \in Q} V_i^q$, donde V_i^q es un dominio (i indica los posibles valores de cada atributo o instancias) de cada uno de los atributos q.
- $f:SxQ \rightarrow V$ es una función de decisión *llamada función de información*, tal que $f(x,q) \in V^q$ para cualquier $q \in Q$, $x \in S$.

El *SI* puede ser representado por una tabla finita de datos, donde las columnas están indicadas por los atributos y las filas por los *objetos*.

Objeto	Atributos						
S	q_1	q_2	...	q_j	...	q_{p-1}	q_p
x_1	V_1^1	V_2^2	V_3^j			V_1^{p-1}	V_1^p
x_2	V_3^1	V_1^2	V_2^j			V_2^{p-1}	V_2^p
. .	V_4^1	V_4^2	V_4^j			V_2^{p-1}	V_k^p

Se denominan *atributos estudiantes* a los atributos comprendidos entre q_1 a q_{p-1}.

Se denomina *atributo experto* o de características al atributo q_p que separa los n objetos en k clases $\{V_1^p, V_2^p, \ldots, V_k^p\}$.

La idea inicial del método de Hunt

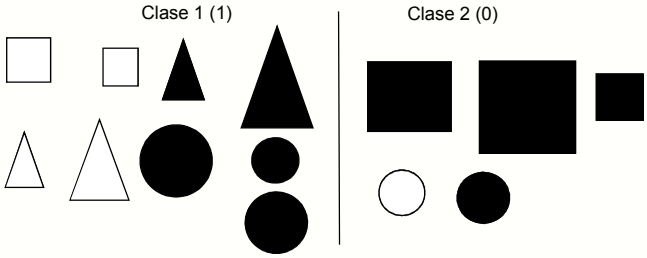
construir un árbol de decisión desde un conjunto de casos de entrenamiento S que consiste de n ejemplos, pertenecientes a k diferentes clases $\{C_1, C_2, \ldots, C_k\}$ indicadas por el atributo experto q_p.

La tarea es dividir el conjunto de entrenamiento S en conjuntos disjuntos $T_1, T_2, \ldots, T_n$, creando una partición, basada en una característica simple.

Ej: Clasificación automática de objetos

Para analizar el caso más general, considere la clasificación de figuras geométricas.

Ej: Clasificación de figuras



La base de datos operacional será:

T	Carac. 1	Carac. 2	Carac. 3	Carac. 4	Carac. 5
Obj.	**Forma**	**Área**	**Tono**	**Sombra**	**Clase**
1	Cuadrado	4 cm^2	Blanco	Si	Clase 1
2	Cuadrado	5 cm^2	Negro	Si	Clase 2
3	Cuadrado	5,5 cm^2	Negro	No	Clase 2
4	Cuadrado	3,9 cm^2	Negro	No	Clase 2
5	Cuadrado	3,8 cm^2	Blanco	No	Clase 1
6	Triángulo	3,9 cm^2	Negro	Si	Clase 1
7	Triángulo	4,6 cm^2	Negro	No	Clase 1
8	Triángulo	3,6 cm^2	Blanco	Si	Clase 1
9	Triángulo	4,2 cm^2	Blanco	No	Clase 1
10	Círculo	3,8 cm^2	Negro	Si	Clase 2
11	Círculo	3,7 cm^2	Blanco	Si	Clase 2
12	Círculo	4 cm^2	Negro	No	Clase 1
13	Círculo	3,7 cm^2	Negro	No	Clase 1
14	Círculo	3,8 cm^2	Negro	No	Clase 1

Particiones

Forma = Cuadrado

Tono = Blanco

	Forma	Área	Tono	Sombra	Clase
1	Cuadrado	4 cm^2	Blanco	Si	Clase 1
5	Cuadrado	3,8 cm^2	Blanco	No	Clase 1

Tono = Negro

2	Cuadrado	5 cm^2	Negro	Si	Clase 2
3	Cuadrado	5,5 cm^2	Negro	No	Clase 2
4	Cuadrado	3,9 cm^2	Negro	No	Clase 2

Particiones

Forma = Triángulo

6	Triángulo	3,9 cm^2	Negro	Si	Clase 1
7	Triángulo	4,6 cm^2	Negro	No	Clase 1
8	Triángulo	3,6 cm^2	Blanco	Si	Clase 1
9	Triángulo	4,2 cm^2	Blanco	No	Clase 1

Particiones

Forma = Círculo

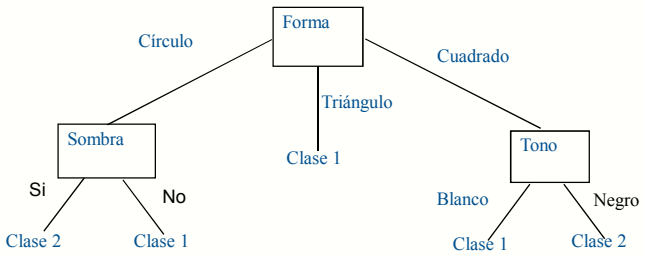
Sombra = Si

10	Círculo	3,8 cm²	Negro	Si	Clase 2
11	Círculo	3,7 cm²	Blanco	Si	Clase 2

Sombra = No

12	Círculo	4 cm²	Negro	No	Clase 1
13	Círculo	3,7 cm²	Negro	No	Clase 1
14	Círculo	3,8 cm²	Negro	No	Clase 1

Árbol



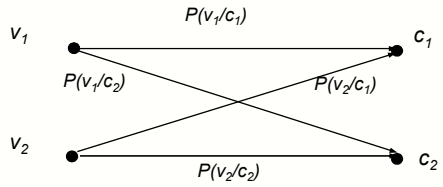
7.2. Cálculo de Entropía y Ganancia de información.

El problema consiste en determinar cual de los atributos estudiantes (V_i $j=1..n-1$), caracteriza de mejor forma las clases C_k.

Considere un problema simple con el factor V_i, que sólo contiene dos instancias ($i=1$e $i=2$) y existen apenas dos clases (C_1 y C_2).

Las relaciones entre las instancias de las características V_i y las clases C_k se pueden relacionar mediante las probabilidades condicionales entre instancias y clases.

Una característica y clase binaria



Si las relaciones se representan en una matriz, se requiere que sólo un elemento de la fila de la matriz sea uno, matrices caso ideal.

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

Se mide la independencia de V_i y C_k mediante probabilidad conjunta $p(v_i \cap c_k)$ o de juntura $p(v_i;c_k)$.

Esto es:

$$\frac{p(v_i;c_k)}{P(v_i)P(c_k)}$$

- Si son independientes la relación será 1.
- Si son completamente dependientes $p(v_i;c_k)=p(v_i)=p(c_k)$ con lo cual la relación será o $1/p(v_i)$ o $1/p(c_k)$.

Para que esta medida sea cero en el caso de independencia, se toma el logaritmo en base 2 de la relación, resultando una medida de información, en bit.

$$ld\left(\frac{p(v_i;c_k)}{P(v_i)P(c_k)}\right)[bit]$$

Esta medida es cero en el caso de ser independientes V y C.

En el caso de ser completamente dependientes la información es: $-ld(p(v_i))$ o $-ld(p(c_k))$.

Para cuantificar la relación de dependencia entre cualquiera de los atributos estudiantes V^j y el atributo experto C, se toma el promedio de la información entre los atributos.

Este promedio se llama: *Ganancia de información*.

$$Ganancia(v^j,c)=\sum_i\sum_k p(v_i;c_k)ld\left(\frac{p(v_i;c_k)}{P(v_i)P(c_k)}\right)$$

Usando la definición de probabilidad condicional: $p(v_i;c_k)=p(c_k/v_i)P(v_i)$ se puede separar en:

$$Ganancia(V,C)=\sum_i\sum_k p(v_i;c_k)ld(p(c_k/v_i))-\sum_i\sum_k p(v_i;c_k)ld(p(c_k))$$

Definiendo

y

$$\alpha=\sum_i\sum_k p(v_i;c_k)ld(p(c_k/v_i))$$

$$\beta=\sum_i\sum_k [p(v_i;c_k)ldP(c_k)]$$

$\beta=\sum_k ldP(c_k)\sum_i p(v_i;c_k)$ **, pero** $\sum_i p(v_i;c_k)=P(c_k)$

así $\sum_k P(c_k)ldP(c_k)$ **= - *información del atributo C*.**

Aplicando la definición de probabilidad condicional en α:

$$\alpha=\sum_i\sum_k p(v_i;c_k)ld(p(c_k/v_i))$$

$$\alpha=\sum_i\sum_k p(c_k/v_i)P(v_i)ld(p(c_k/v_i))$$

$$\alpha=\sum_i P(v_i)\sum_k p(c_k/v_i)ld(p(c_k/v_i))$$

Definiendo $inf(C/v_i) = -\sum_k p(c_k / v_i) ld(p(c_k / v_i))$, como la información de la clase C condicionada (particionada) por la instancia v_i del atributo estudiante V^j.

El promedio de la información de C condicionada por V^j será la ponderación de la información particionada:

$$\boldsymbol{Inf(C/V)} = \sum_i P(v_i)\, inf(C / v_i)$$

La ganancia (de información) será:

$$Ganancia(V) = Inf(C) - Inf(C/V)$$

Realizando una analogía con el canal de comunicación se tiene que:

Ganancia = I : Información Mutua del Canal
Inf(C)=H(C) entropía del receptor
Inf(C/V) =H(C/V) entropía de error en la recepción
I=H(C) - H(C/V).

Con esta ganancia es posible determinar cual de los atributos estudiantes V^j ($j=1..p-1$) separa o caracteriza de una forma más adecuada las clases c_k.

Para realizar esto se calcula:

$$\max_{j} (Ganancia(V^j))$$

El atributo j será la raíz del árbol.

Para el ejemplo anterior:

Ganancia(Forma)= Inf(Clase)-Inf(Clase/Forma)

Inf(Clase)= -9/14 ld(9/14) - 5/14 ld(5/14) = 0,94 bit.

Inf(Clase/Forma)= 5/14 (-2/5 ld(2/5) - 3/5 ld(3/5))
+ 4/14 (-4/4 ld(4/4) - 0/4 ld(0/4))
+ 5/14 (-3/5 ld(3/5) - 2/5 ld(2/5))
= 0,694 bit.

Ganancia(Forma)= 0,94 - 0,694 = 0,246

De la misma forma se puede calcular la ganancia para los demás atributos estudiantes:

Ganancia(Sombra)= 0,94 - 0,892 = 0,048 bit

Ganancia(Tono) = 0,94 - 0,8949 = 0,045 bit

Ganancia(Área) = 0,94 - 0,9371 = 0,0029 bit.

- Modificación de Ganancia

El criterio de ganancia produce buenos resultados para atributos con cantidades similares números de instancias.

Produce graves distorsiones cuando existen atributos con diferente numero de instancias.

Si un atributo V_i^j $(i=1,2,3...n)$ tiene muchas instancias la información condicional Inf(C/V), disminuye, aumentando artificialmente la ganancia.

La ganancia aumenta puesto que:

Ganancia(V)= Inf(C)-Inf(C/V)

Pero al aumentar las instancias también aumenta la información contenida en el atributo (Split information").

Esto se utiliza para normalizar la Ganancia y generar una nueva relación denominada Razón de Ganancia.

Así: *Split Inf(V) = Inf(V) = H(V)=* $-\sum_i P(v_i)ld(P(v_i))$

Razón Ganancia(V) = Ganancia(V)/Inf(V)

Esta normalización de la ganancia permite eliminar el sesgo introducido por la medida de información que depende del número de instancias.

Este factor de corrección puede cambiar la decisión respecto a los atributos que están mas cercanos a la raíz del árbol.

Para el ejemplo anterior:

Split Inf(V) = -5/14 ld(5/14) - 4/14 ld(4/14) - 5/14 ld(5/14)
= 1,577 bit.

la Razón de Ganancia(V) = 0,246/1,577 = 0,156

Razón Ganancia(Sombra)= 0,048/0,98523 = 0,04872

Razón Ganancia(Tono) = 0,045/0,93977 = 0,04788

Razón Ganancia(Área) = 0,0029/0,93977 = 0,0031.

Para este caso en particular se mantiene la relación de importancia de los atributos estudiantes.

7.3. Poda en árboles de decisión

El método de partición recursiva, continúa subdividiendo los casos de entrenamiento hasta que cada sub-conjunto contenga **casos de una sola clase** o hasta que no existan mas atributos estudiantes para dividir

El resultado de este proceso es un árbol muy complejo y muchas veces **sobre ajustado** a los datos de entrenamiento.

La idea básica consiste en remplazar una parte del árbol por una **hoja simple** que tenga como clase representante la **clase de mayor frecuencia**

En general existen muchas estrategias de poda.

Puede ser ***pre-poda*** si se realiza en la etapa de construcción del árbol o ***poda*** si se realiza retrospectivamente.

El mecanismo sugerido por Quinlan es la poda, permite comparar el beneficio de la poda en relación al árbol original (árbol sobre ajustado).

En el caso de la poda, el árbol de decisión es simplificado descartando uno o más sub-árboles y reemplazándolo por hojas.

La clase de la hoja se encuentra buscando los casos de entrenamiento que se asocian mayoritariamente a una clase.

En cada hoja del árbol entrenado se muestra la clase a la cual se asocia la hoja y la relación (N/E)
N: indica él número total de casos asignados a la hoja
E: número de casos mal clasificados en la clase indicada.

Ej: ***Clasificación de paciente según los requerimientos de enfermería.***

Atributos estudiantes:

Tipo de cirugía	***{leve, mediana, compleja}***
Ambulación	***{ayuda, Independiente, camilla}***
Dependencia	***{auto cuidado, moderado, Intensivo}***
Grado de invasión	***{leve, moderado, grande}***
Estado psicológico	***{tranquilo, irritado, agresivo}***
Estado cognitivo	***{orientado, desorientado, inconsciente}***
Edad	***{<50, 50-70, >70}***

Atributo experto: 1: baja demanda, 2: alta demanda.

Ej: Requerimientos de enfermería Árbol Entrenado.

Tipo de cirugía = leve
- *Grado invasión =leve: 1 (151)*
- *Grado invasión =moderado: 1 (1)*
- *Grado invasión =grande*
  - *Edad =<50: 1 (6)*
  - *Edad =50-70: 1 (9)*
  - *Edad =≥70: 2 (1)*

Tipo de cirugía = mediana
- *Ambulación = camilla: 2 (97/3)*
- *Ambulación = ayuda: 2 (4)*
- *Ambulación = independiente*
  - *Estado cognitivo = orientado: 1 (2)*
  - *Estado cognitivo = desorientado: 2 (1)*
  - *Estado cognitivo = inconsciente*
    - *Edad =<50: 1 (5/2)*
    - *Edad =50-70: 2 (13/2)*
    - *Edad =≥70: 1 (1)*

Ej: Requerimientos de enfermería Primera Rama.

El primer sub-árbol

Tipo de cirugía = *leve*

Grado invasión =leve: 1 (151)

Grado invasión =moderado: 1 (1)

Grado invasión =grande

Edad =<50: 1 (6)

Edad =50-70: 1 (9)

Edad =≥70: 2 (1)

Si se reemplaza esta rama por Tipo de cirugía = leve: 1 (168/1), se cometerá un error de 1 caso al considerarlo como Tipo 1 cuando es Tipo 2.

Ej: Segunda rama del árbol

Tipo de cirugía = mediana

Ambulación = camilla: 2 (97/3)

Ambulación = ayuda: 2 (4)

Ambulación = independiente

Estado cognitivo = orientado: 1 (2)

Estado cognitivo = desorientado: 2 (1)

Estado cognitivo = inconsciente

Edad =<50: 1 (5/2) (3)

Edad =50-70: 2 (13/2)

Edad =≥70: 1 (1)

Si se reemplaza toda la rama por la hoja Tipo de cirugía = mediana: 2 (123/11), se tendrán 11 casos mal clasificados en contraste a los 7 mal clasificados del árbol completo.

Se puede estimar el error en la ***población*** con el árbol sin podar y podado (hoja).

Se usa una distribución de probabilidad conocida con un cierto límite de confianza.

Hipótesis: La probabilidad de ocurrir E errores en *N* ensayos esta dada por una distribución Binomial de probabilidad p en la población.

B(r,n,p)

E=r el número de errores
N=n el número de ensayos y
p: probabilidad de error esperada en la población.

La densidad está dada por:

$$f(r)=\binom{n}{r}p^n(1-p)^r$$

Dados ***n*** y ***r***, se requiere determinar ***p*** para un cierto nivel de confianza ***Co%.***

$$\text{Con Co}=\sum_{i=0}^{r} f(i)$$

esto será $P_{Co}(r,n)$ o $P_{Co}(E,N)$.

Entonces, el número de errores que se tendrá en la población, para una hoja que en el modelo clasifica ***N*** casos con ***E*** errores será:

$N \times P_{Co}(E,N)$.

Quinlan usa un nivel de confianza que denomina pesimista al *50%*, pero como la distribución Binomial es simétrica, basta con usar ***Co/2***, esto es un nivel del **25%** unilateral.

Para el ejemplo anterior se tiene el siguiente sub-árbol:

Edad =<50: 1 (6)
Edad =50-70: 1 (9)
Edad =≥70: 2 (1)

La primera hoja: *N=6*, *E=0*, de las tablas de la distribución Binomial $P_{25\%}(0,6)=0,206$.

Si la hoja fuera usada para predecir 6 casos desconocidos, el error sería: *6x0,206=1,236.*

Para la segunda hoja: $P_{25\%}(0,9)=0,143$. *Error población 9x0,143=1,287.*

Para la tercera hoja:

$P_{25\%}(0,1)=0,750$. *Error población 1x0,75=0,75.*

El error para todo el sub-árbol será:

1,236+1,287+0,75=3,273.

Si este sub-árbol fuera reemplazado por una hoja correspondiente a la clase Tipo 1 (la más frecuente), se cubrirían los mismos 16 casos con un error de un caso.

Su error de predicción sería:

$$P_{25\%}(1,16) = 0,157$$

El error en la población será:

$$16x0,157 = 2,512.$$

Dado que la hoja presenta un error inferior al sub-árbol debe ser ***reemplazado*** por la hoja de Tipo 1.

Sustituyendo esta hoja, el árbol superior queda:

Grado invasión =leve: 1 (151)
Grado invasión =moderado: 1 (1)
Grado invasión =grande: 1 (16/1)

El número de errores predichos para este árbol será:

$$151xP_{25\%}(0,151)+1xP_{25\%}(0,1)+2,512=4,642.$$

Si se desea reemplazar este sub-árbol por una hoja que clasifique en el Tipo 1, el error predicho para esta hoja será: ***$168xP_{25\%}(1,168)=2,610.$***

Lo cual es inferior al valor del sub-árbol.

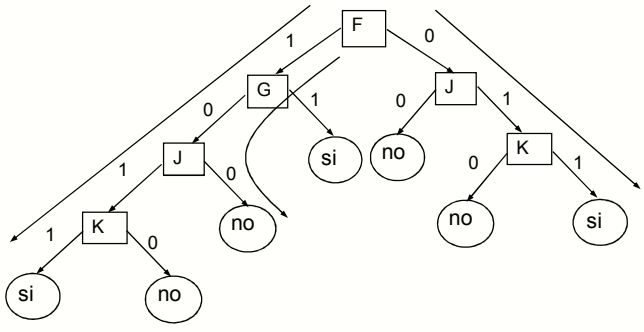
Por lo tanto, el árbol puede ser podado por la hoja correspondiente.

7.4. Transformando árboles en reglas

Suponga el siguiente árbol genérico

```
F = 0
|   J = 0:no
|   J = 1
|   |   K = 0: no
|   |   K = 1: si
F = 1
|   G = 1: si
|   G = 0
|   |   J = 0: no
|   |   J = 1:
|   |   |   K=0: no
|   |   |   K=1: si
```

Definición: Una ***regla*** corresponde a un ***camino*** entre la ***raíz*** y cada una de las ***hojas***.



El transcurso entre la raíz y la hoja corresponde a la condición de la regla, denominado ***antecedente***.

El valor de la hoja es la conclusión de la regla denominado ***consecuente***.

Si *(F=0 ^ J=1 ^ K=1)* ***Entonces*** **si**
(F=0 ^ J=1 ^ K=1) Antecedente; **si** Consecuente

Al tratar de representar todas las posibles reglas que se representan en un árbol se puede tener una estructura más compleja que el propio árbol.

Sin embargo es posible observar que el antecedente de una regla en particular puede contener condiciones irrelevantes.

Si se examina la rama derecha y la izquierda si tienen las siguientes reglas

Si *(F=0∧J=1∧K=1)* ***entonces*** la clase es (si)
Si *(F=1∧G=1)* ***entonces*** la clase es (si)
Si *(F=1∧G=0∧J=1∧K=1)* ***entonces*** la clase es (si)

Esta regla puede ser generalizada como:

Si (J=1∧K=1) ***entonces*** la clase es (si).

árboles generan reglas redundantes.

¿como se pueden eliminar las condiciones irrelevantes?

- Reducción de reglas

Sea R una regla que contiene en su antecedente (A).

R: Si (A) entonces clase ***(c)***

Si se le elimina la condición A_i.

Se tiene una regla (R-) Especializada

R-: Si (A-) entonces clase ***(c)***

La evidencia de la importancia de la condición A_i debe ser encontrada en los casos de entrenamiento.

Cada caso en el antecedente generalizado ***A-*** puede pertenecer o no pertenecer a la clase ***C***

Por otro lado puede satisfacer o no satisfacer la condición A_i.

Esto genera cuatro grupos, que son organizados en una tabla de contingencia.

	Clase *C* ***(casos de A-)***	**Otras clase (Casos de A-)**
Satisfacen A_i	X_P	E_P
No satisfacen A_i	X_N	E_N

Los casos positivos (X_P+E_P) que satisfacen $A-$ y A_i, son cubiertos por la regla original R, y los E_P son mal clasificados por la regla R.

Los casos negativos (X_N+E_N) que satisfacen $A-$, pero no satisfacen A_i, pueden ser cubiertos por la regla generalizada $R-$ pero no por la regla original R. Existen E_N errores de clasificación.

Como la regla generalizada $R-$ cubre todos los casos que satisfacen a la regla especializada R, el total de casos cubiertos por R es $(X_P+E_P+X_N+E)$.

Quinlan propone usar $P_{Co}(E,N)$. Con nivel Co.

-Para la regla especializada R se tiene: $P_{Co}(E_P, X_P+E_P)$.

- Para la regla generaliza $R-$ se tiene:

$$P-_{Co}(E_P+E_N, X_P+E_P+X_N+E_N).$$

Si $P-_{Co} > P_{Co}$ se elimina A_i.

Permite eliminar una condición A_i de un conjunto de antecedentes A.

Si se ordenan todas las posibles condiciones y sus respectivos errores estimados, es posible eliminar primero el menor error estimado, para eliminar la condición que aporta el menor error estimado.

Luego en una etapa siguiente se calcula nueva mente los errores estimados para cada una de las condiciones restantes y se elimina nuevamente la condición con menor error estimado, y así sucesivamente.