

APRENDIZAJE IINDUCTIVO MEDIANTE ÁRBOLES DE DECISON

INDICE

- 1.- Ideas Generales sobre árboles de decisión.**
- 2.- La familia TDIDT. Un algoritmo genérico de aprendizaje**
- 3.- El algoritmo ID3. Problemas en la aplicación de ID3**
- 4.- Variantes del algoritmo ID3.**
- 5.- Metodología CART**
- 5.- Discusión**

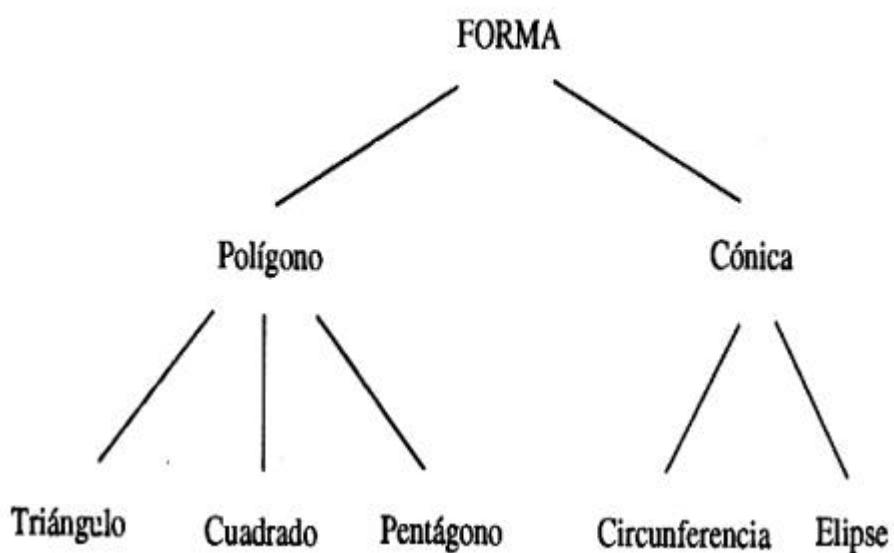
IDEAS GENERALES SOBRE ÁRBOLES DE DECISION

Hablando de un modo muy general, un árbol de decisión es un árbol en el que:

- Los nodos representan conjuntos de objetos,
- Los arcos que parten de un nodo representan un criterio de partición del conjunto asociado al nodo del que parten.

El nombre alude que un árbol de decisión siempre representa un proceso de decisión relativo a los objetos de un conjunto (que será el nodo raíz) y viceversa.

Ejemplo:



Los árboles de decisión son herramientas especialmente idóneas para representar los procesos de decisión que aparecen involucrados en cualquier proceso inductivo de clasificación.

En la literatura se han desarrollado técnicas para construir un árbol de decisión a partir de un conjunto de ejemplos de una clasificación, árbol que encapsula el conocimiento contenido en el citado conjunto de ejemplos.

Estas técnicas han recibido el nombre de TDIDT (Top Down Induction of Decision Trees) debido a la forma en que construyen dicho árbol.

Los métodos incluidos dentro de esta familia emplean como sesgo inductivo la idea e obtener el árbol más pequeño posible que clasifica todos los ejemplos.

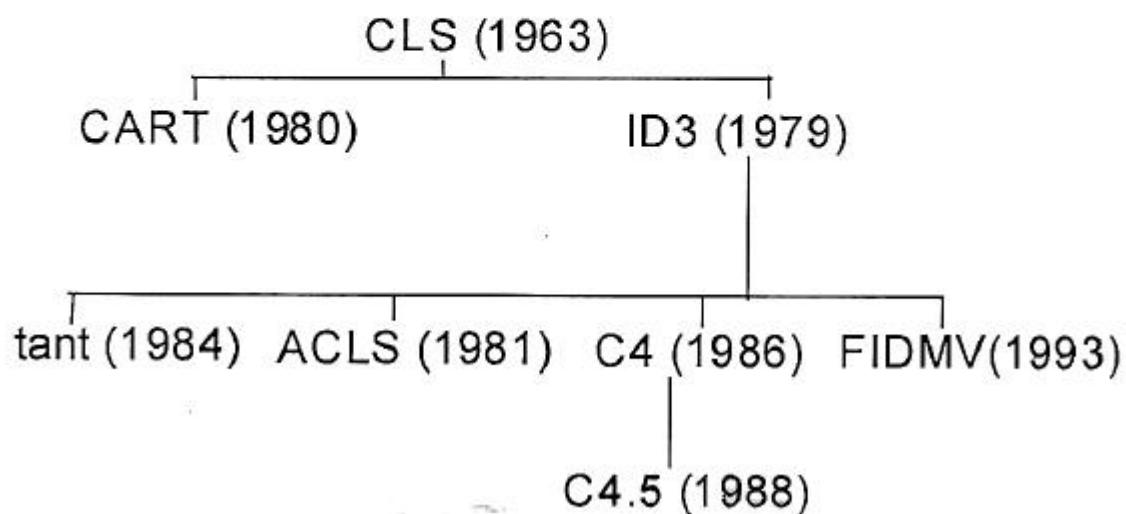
En lo que difieren unos métodos y otros es en el criterio para llegar a ese árbol minimal.

¿Porqué favorecer las hipótesis más cortas?

- William de Occam (1320)
 - > Occam's razor (la navaja de Occam)

- es más difícil encontrar una hipótesis simple que describa un fenomeno que una hipótesis compleja.
- las hipótesis simples son hipótesis más generales

Algoritmos de aprendizaje con árboles



El algoritmo CLS (Concept Learning System) es el padre de todo la familia y trataba de obtener un árbol que clasificase con el menor costo (computacional).

El más famoso de toda la familia es ID3 (Interactive Dichotomizer 3), introducido por Quinlan empleando un criterio de minimización basado en criterios de “ganancia de información”.

Nosotros nos vamos a centrar en ID3.

ALGORITMO GENERICO

$$E = \{x, c(x), x \in X\} \quad x = (x_1, \dots, x_n)$$
$$x_i : a_i : s_i$$

- 0.- Conjunto de hojas $H = \{E\}$
- 1.- Escoger HOJA $\in H$. Borrar HOJA de H
- 2.- Escoger un atributo a_i
- 3.- Construir una pregunta P sobre a_i de respuestas R_1, R_2, \dots, R_w
- 4.- Obtener H_1, H_2, \dots, H_w como resultado de aplicar P a HOJA. Añadirlas a H .
- 5.- Si se cumple
 $\forall H \in H, \forall x, x' \in H \quad c(x) = c(x')$
entonces stop.

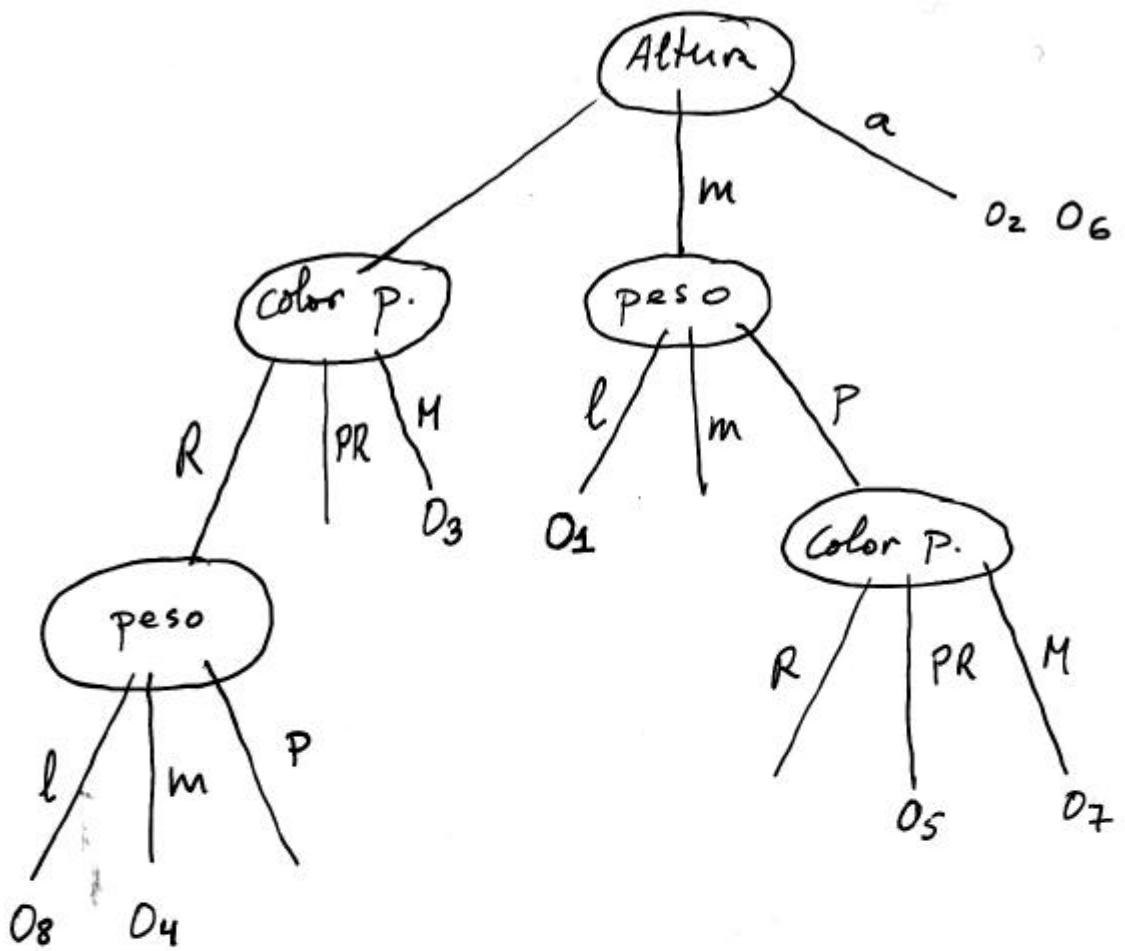
En caso contrario determinar un \hat{H} para el que no se cumpla. Hacer HOJA = \hat{H} y volver a 2.

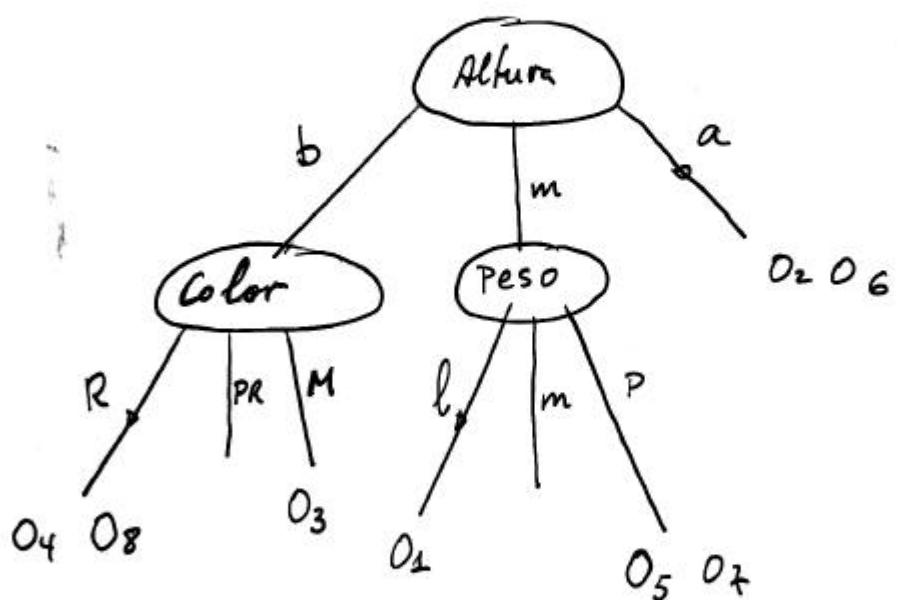
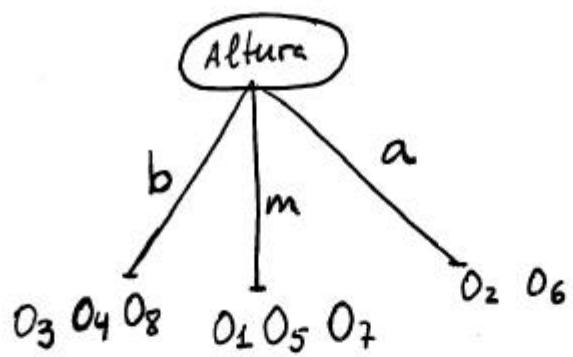
CUESTIONES

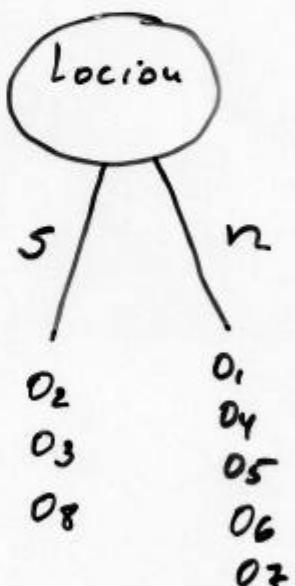
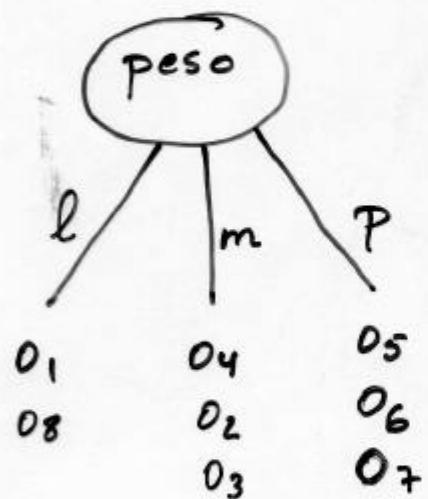
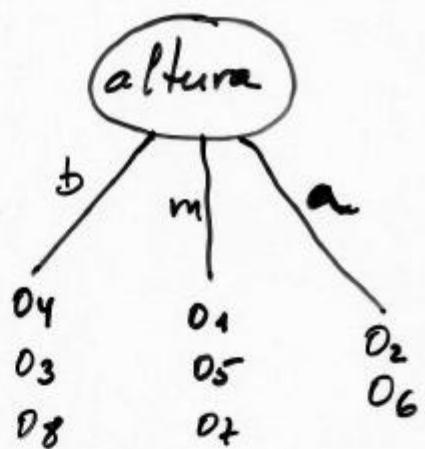
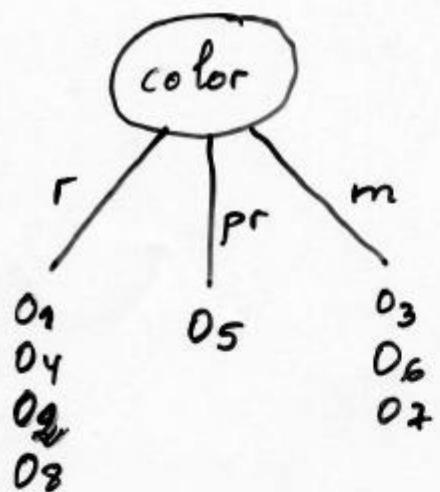
- Tipo de función c : binario, no binario, simbólico.
- Como escoger \hat{H} , como escoger a_i
- Como diseñar la pregunta P .

EJEMPLO

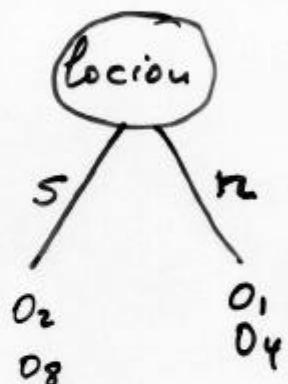
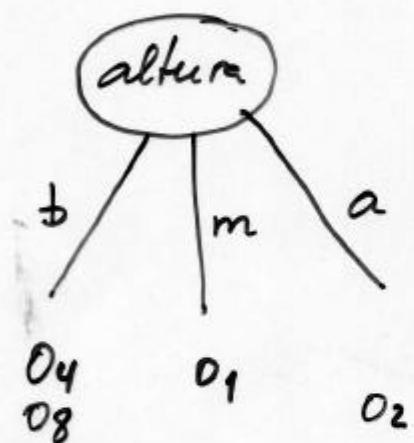
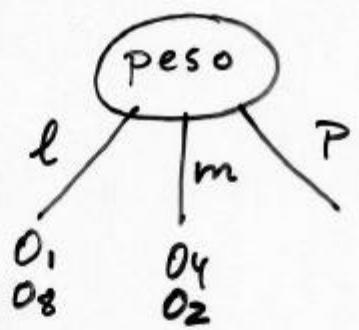
Nombre	Pelo	Estatura	Peso	Loción	Resultado
O ₁	Rubio	Medio	ligero	no	quemado
O ₂	R	Alto	Medio	si	nada
O ₃	Moreno	bajo	medio	si	nada
O ₄	R	bajo	medio	no	quemado
O ₅	P.Rojo	medio	pesado	no	quemado
O ₆	M	Alto	pesado	no	nada
O ₇	M	Medio	pesado	no	nada
O ₈	R	Bajo	ligero	si	nada



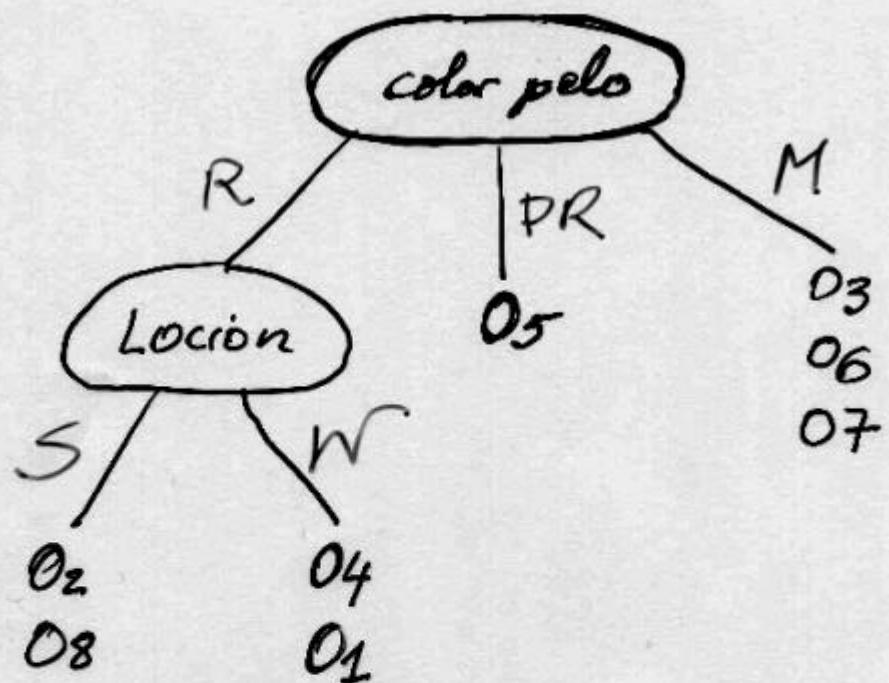




Minimizar desorden : color de pelo



Minimizar desorden : locion



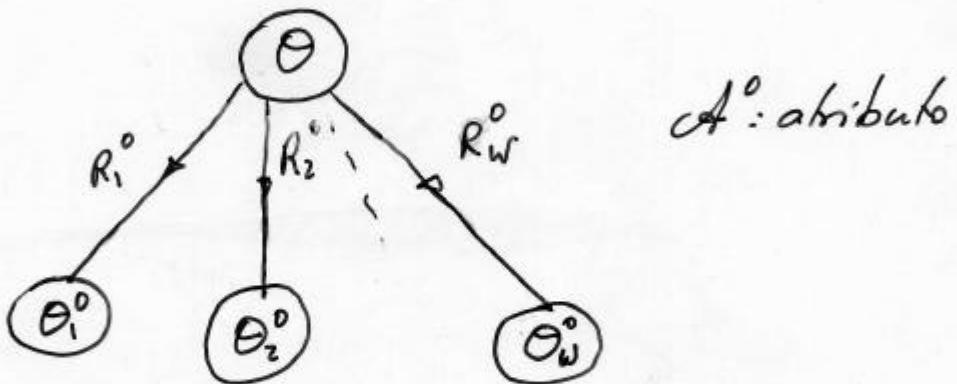
ID3 (Interactive Dichotomizer 3)

Quinlan (1970-1976)

$$\Theta = \{ [x, c(x)] , x \in X_\Theta \} \quad N = \text{card}(\Theta)$$

$$c(x) \in \{ C_1, \dots, C_M \}; \quad X_\Theta = \bigcup_i^M X_i, \quad N_i = \text{card}(X_i)$$

$$\text{proto}(x \in X_i) = \frac{N_i}{N}$$



$\Theta_j^\circ = \{ \text{objetos de } \Theta \text{ con respuesta } R_j^\circ \text{ a la pregunta sobre el atributo } A^\circ \}$

$$N_j^\circ = \text{card}(\Theta_j^\circ)$$

$N_{j,i}^\circ = \text{card}(\Theta_j^\circ \cap X_i)$: número de elementos de la clase i en Θ que tienen respuesta R_j° a la pregunta sobre A°

$$H(\Theta_j^o) = - \sum_{i=1}^M \frac{N_{ji}^o}{N_j^o} \lg_2 \frac{N_{ji}^o}{N_j^o}$$

$$E(A^o) = \sum_{j=1}^W \frac{N_j^o}{N} H(\Theta_j^o)$$

$$g(A^o) = H(\Theta) - E(A^o)$$

ID3: Escoger en cada nodo el atributo A^o que maximiza la ganancia de información en relación con la clase que se divide

$$\text{Max } g(A^o) \Leftrightarrow \text{Min } E(A^o)$$



ID3: criterio de mínima entropía

NOTA Empleo efectivo de solo una parte del conjunto de entrenamiento (VENTANA). El resto se debe emplear como conjunto de test.

EJEMPLO ID 3

$$\bar{X} = \{O_1, O_2, \dots, O_8\} \quad C = \left\{ \begin{matrix} C_1 & C_2 \\ Q & NQ \end{matrix} \right\}$$

$$\bar{X}_1 = \{O_1, O_4, O_5\}; \quad \bar{X}_2 = \{O_2, O_3, O_6, O_7, O_8\}$$

PRIMERA ITERACIÓN

$$\Theta = \bar{X} = \{O_1, O_2, O_3, O_4, O_5, O_6, O_7, O_8\}$$

Color de pelo $\{R, PR, M\}$

$$\Theta_R = \{O_1, O_4, O_2, O_8\} \quad \frac{4}{8} \quad \frac{Q}{Z} \quad \frac{NQ}{Z}$$

$$\Theta_{PR} = \{O_5\} \quad \frac{1}{8} \quad 1 \quad 0$$

$$\Theta_M = \{O_3, O_6, O_7\} \quad \frac{3}{8} \quad 0 \quad 3$$

$$E(\text{color pelo}) = \frac{4}{8} \left(-\frac{2}{4} \log_2 \frac{2}{4} - \frac{2}{4} \cdot \log_2 \frac{2}{4} \right)$$

$$+ \frac{1}{8} \cdot 0 + \frac{3}{8} \cdot 0 = 0.5$$

$$E(\text{peso}) = 0.94$$

$$E(\text{talla}) = 0.69$$

$$E(\text{Loción}) = 0.61 \quad (*)$$

} Mínimo

Nueva clase

$$\Theta = \{0, 0_4, 0_2, 0_8\}$$

Atributo: Altura = {B, M, A}

$$\Theta_B = \{0_4, 0_8\} \quad \frac{2}{4} \quad \frac{\alpha}{1/2} \quad \frac{N\alpha}{1/2}$$

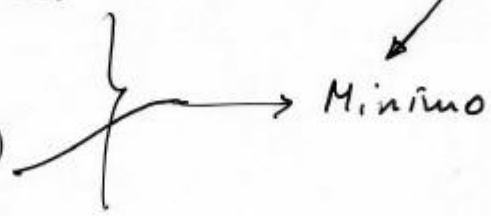
$$\Theta_M = \{0_1\} \quad \frac{1}{4} \quad 1 \quad 0$$

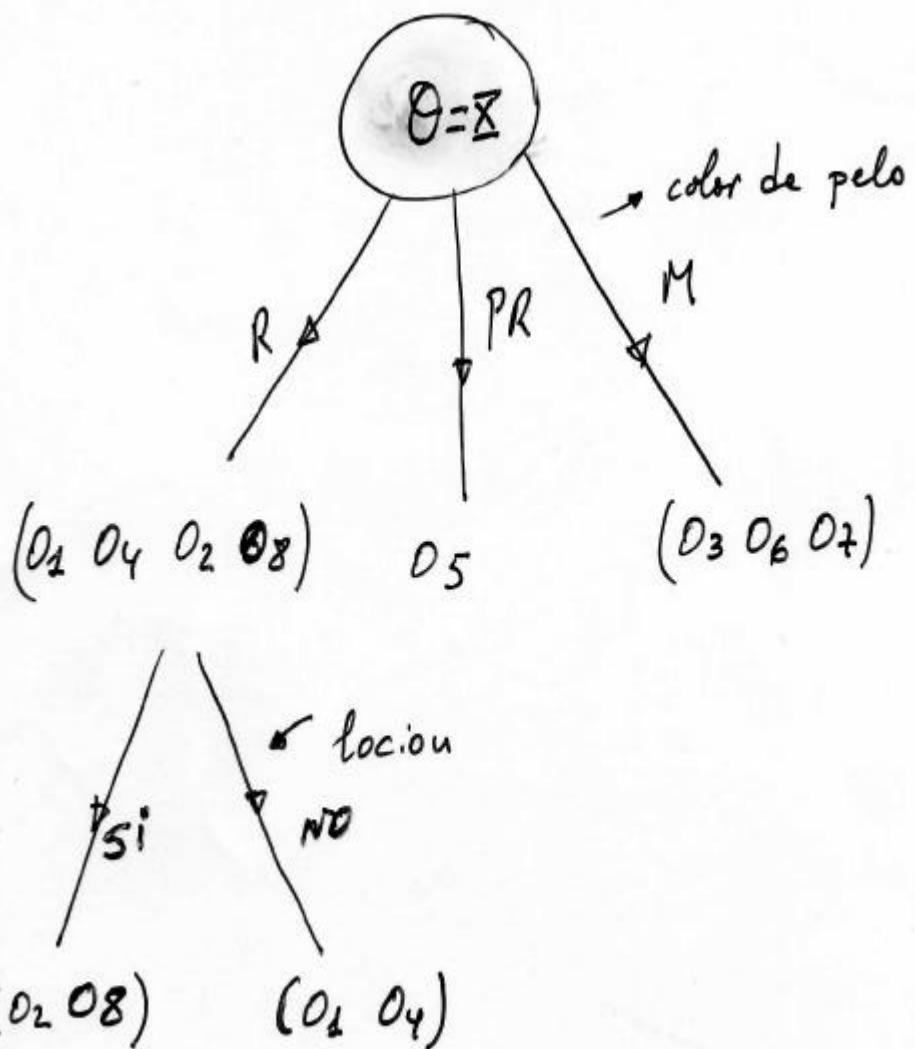
$$\Theta_A = \{0_2\} \quad \frac{1}{4} \quad 0 \quad 1$$

$$E(\text{Altura}) = \frac{2}{4} \left(-\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2} \right) \\ + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0 = 0.5$$

$$E(\text{peso}) = 1$$

$$E(\text{loc.ou}) = 0$$





DESCRIPCIÓN MEDIANTE REGLAS

IDEA GENERAL: Cada camino del nodo raíz a una hoja describe una regla SI-ENTONCES que determina el concepto.

Ejemplo quemados

Si : la persona es rubia
la persona usa loción } R₁
Entonces: no pasa nada

Si la persona es morena
la persona no usa locón } R₂
Entonces se quema

Si la persona es pelirroja
Entonces x quema } R₃

Si la persona es morena
Entonces no pasa nada } R₄

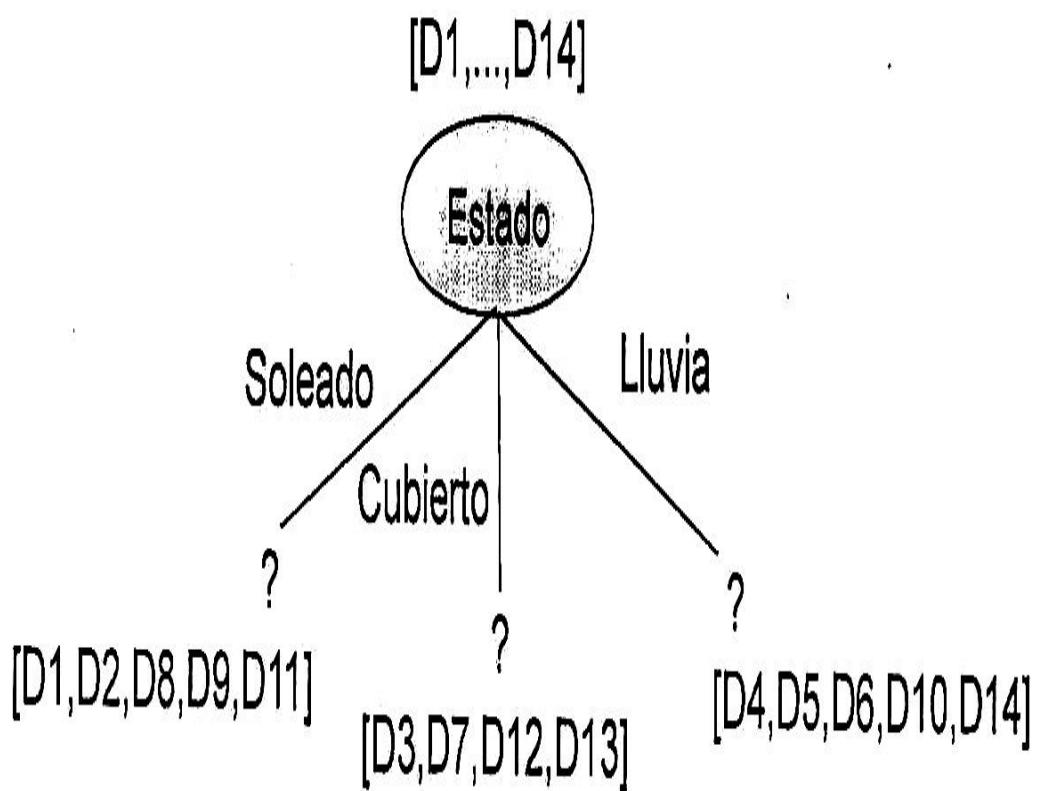
SIMPLIFICACIÓN DE REGLAS

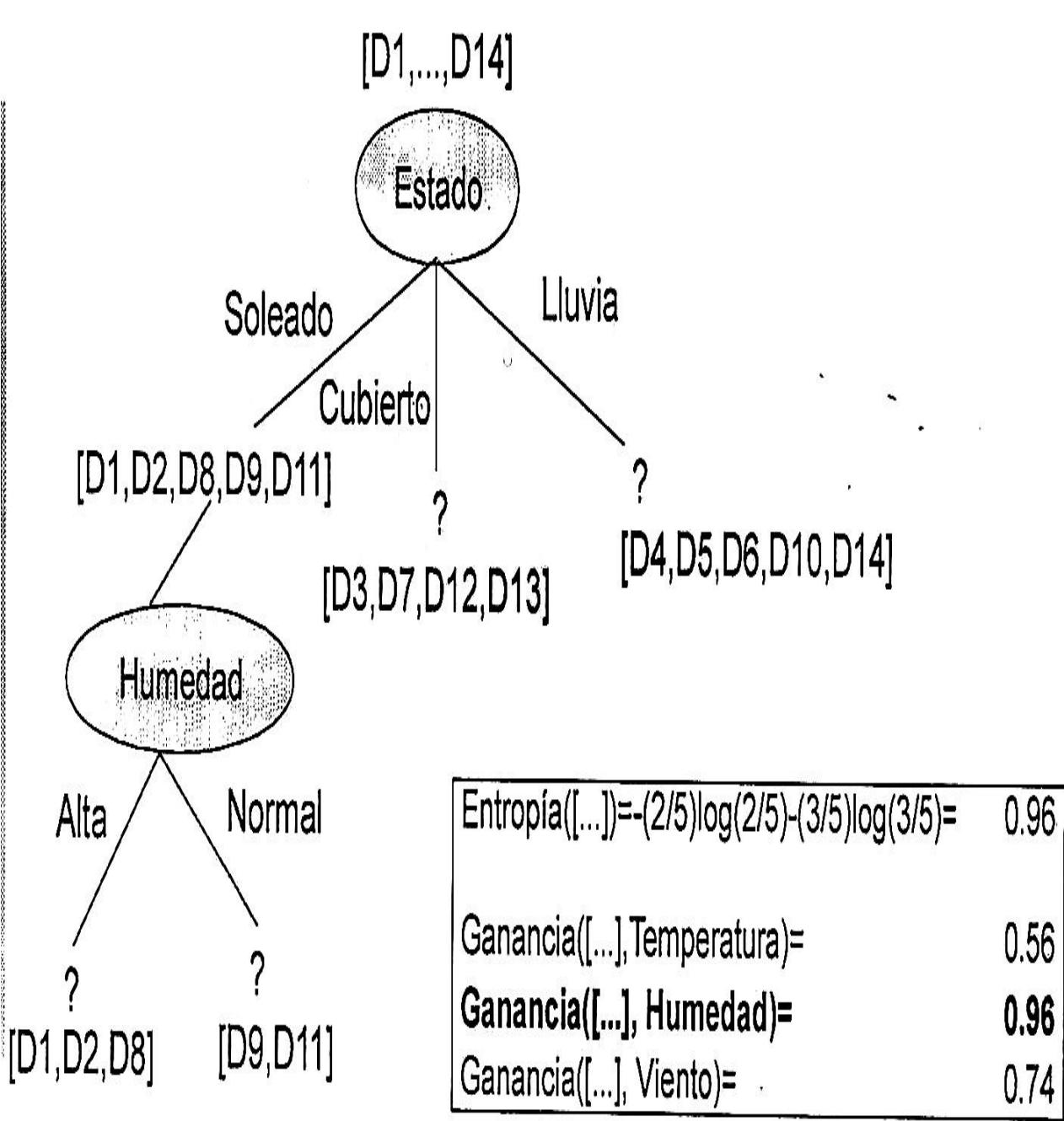
OTRO EJEMPLO

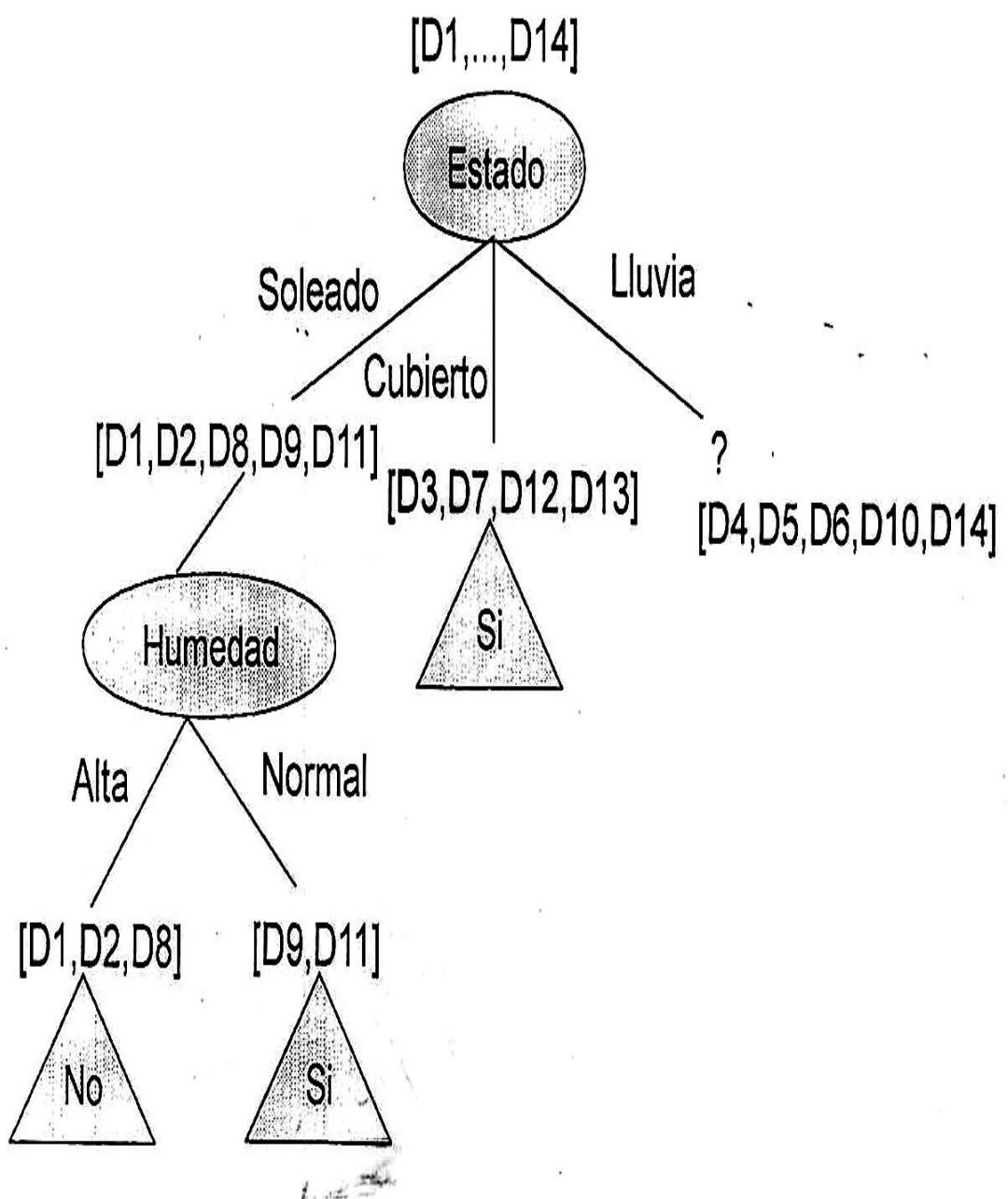
Día	Estado	Temperatura	Humedad	Viento	JugarTenis
D1	Soleado	Calor	Alta	Débil	No
D2	Soleado	Calor	Alta	Fuerte	No
D3	Cubierto	Calor	Alta	Débil	Si
D4	Lluvia	Agradable	Alta	Débil	Si
D5	Lluvia	Frío	Normal	Débil	Si
D6	Lluvia	Frío	Normal	Fuerte	No
D7	Cubierto	Frío	Normal	Fuerte	Si
D8	Soleado	Agradable	Alta	Débil	No
D9	Soleado	Frío	Normal	Débil	Si
D10	Lluvia	Agradable	Normal	Débil	Si
D11	Soleado	Agradable	Normal	Fuerte	Si
D12	Cubierto	Agradable	Alta	Fuerte	Si
D13	Cubierto	Calor	Normal	Débil	Si
D14	Lluvia	Agradable	Alta	Fuerte	No

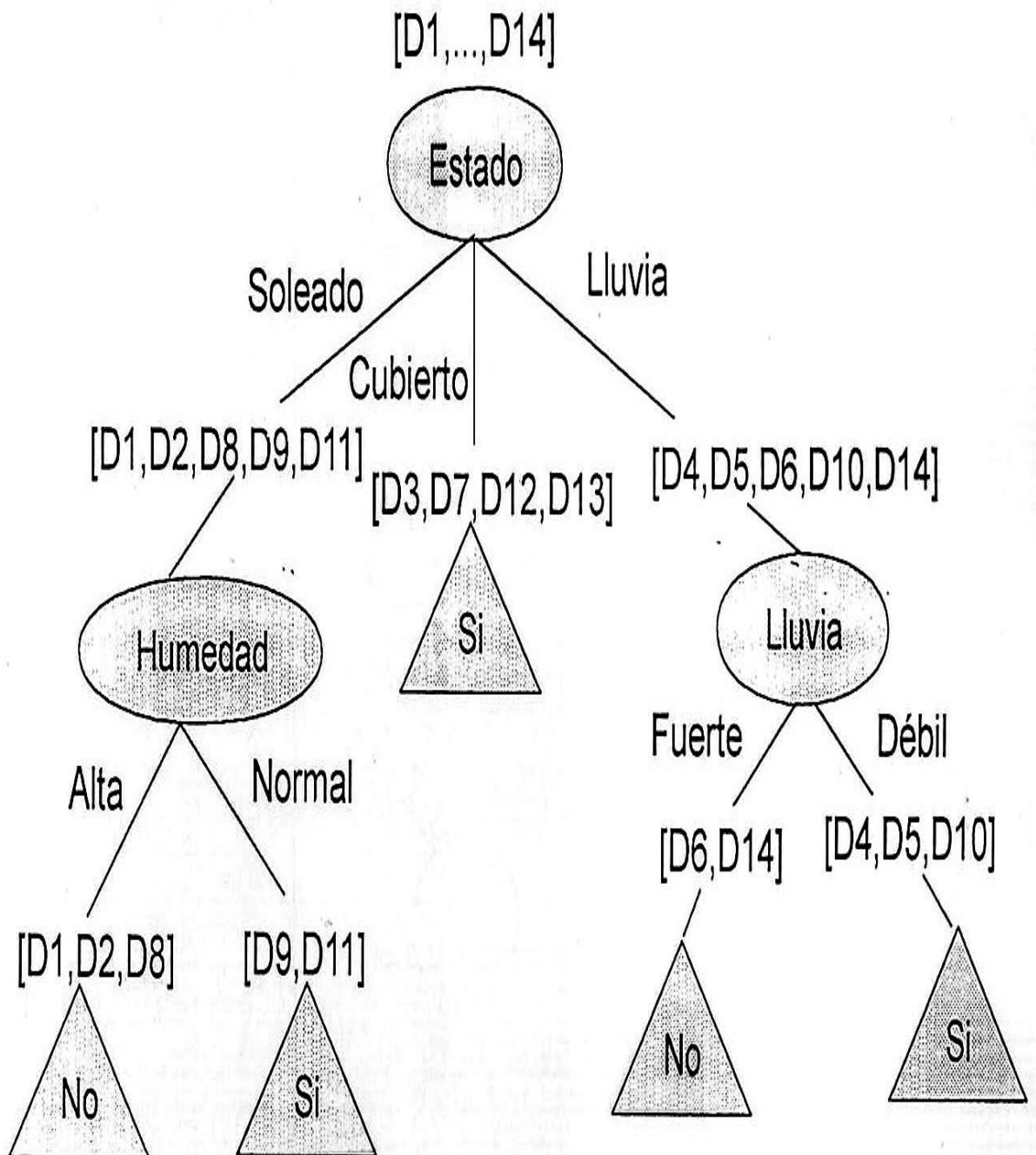
Tabla 1: Conjunto de ejemplos

Ganancia (E,Estado)	0.891
Ganancia (E,Temperatura)	0.029
Ganancia (E,Humedad)	0.151
Ganancia (E,Viento)	0.048









5. Algunos aspectos sobre el aprendizaje en A.D.

ID3 presenta algunas deficiencias:

- problemas a los que se puede aplicar
- características del conjunto de ejemplos (ruido)

Sus descendientes corren estas deficiencias:

- * el sobre-ajuste
- * atributos con valores continuos
- * otras medidas para seleccionar atributos
- * atributos con valores perdidos

5.1 El problema del sobre-ajuste

ID3 ramifica hasta que todos los ejemplos quedan correctamente clasificados.

Si el conjunto está afectado de ruido
ID3 intenta clasificar también el ruido

Definición: Decimos que una hipótesis sobre-ajusta el conjunto de entrenamiento si existe otra hipótesis que describe menos consistentemente a los ejemplos, pero un mejor comportamiento sobre la generalización de los mismos.

5.1 El problema del sobre-ajuste (2)

Ejemplo:

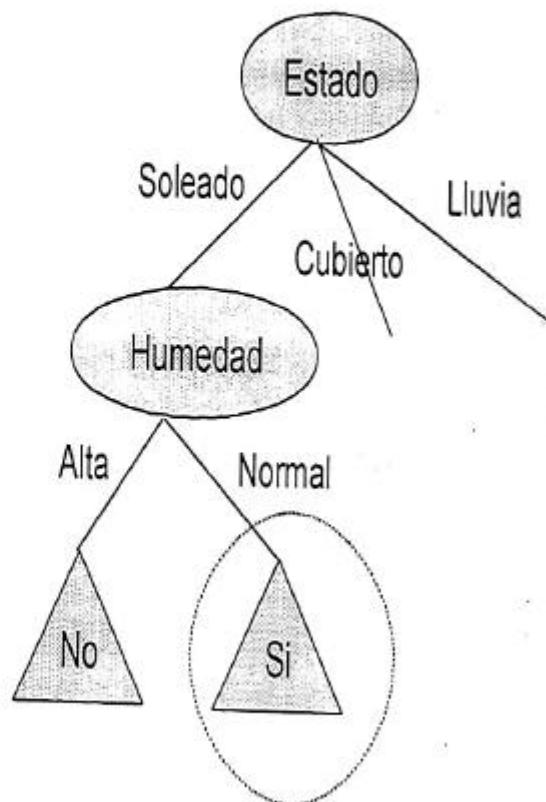
Estado = Soleado

Temperatura=Calor

Humedad=Normal

Viento=Fuerte

JugarTenis=No



5.1 El problema del sobre-ajuste (3)

El sobre-ajuste es un problema importante.

En A.D. se han propuesto 2 vias:

- deterner el crecimiento del árbol durante la fase de creación (pre-poda)
- un segundo proceso se encarga de reducir el sobre-ajuste con una poda (post-poda).

5.1. El problema del sobre-ajuste (4)

Se han estudiado algunas vías para estimar el tamaño correcto del árbol:

- Dividir el conjunto de ejemplos disponibles en: conj. entrenamiento y conj. validación.
- Usar todos los ejemplos disponibles y usar un test estadístico para estimar si hay que expandir un nodo. (Quinlan Chi-Cuadrado)

5.1.1. Poda para reducir el error

Considera todos los nodos no hoja como candidatos para ser podados:

- * si se poda se sustituye por la clase predominante
- * se poda si su eliminación no empeora la capacidad de predicción del árbol resultante sobre el conj. validación.
- * es iterativo empezando por los nodos que provocan una mejora en la clasificación.

5.2. La regla de post-poda

¿Qué ocurre si hay pocos ejemplos disponibles?

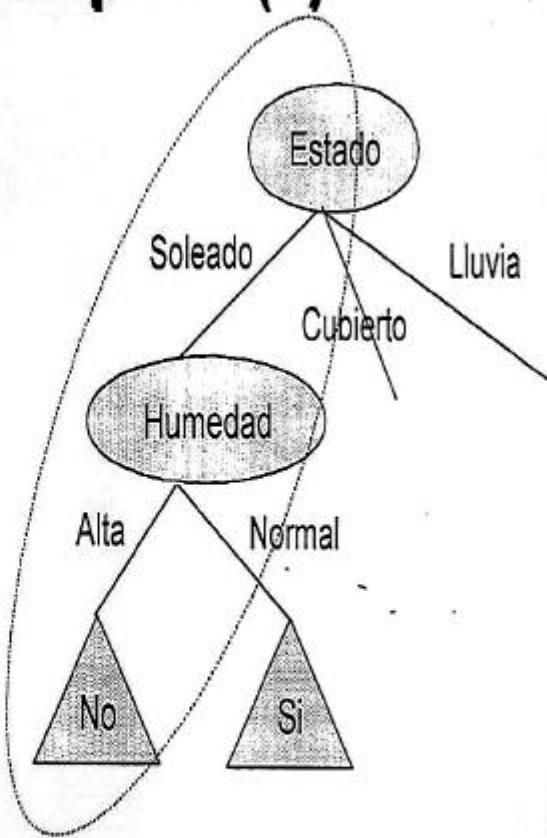
PROCESO

1. Construir el árbol
2. Transformar el árbol en reglas
3. Podar cada regla (eliminar un atributo)
4. Ordenar las reglas por cap. de predicción

5.2. La regla de post-poda (2)

Ejemplo:

IF (Estado=Soleado)
and (Humedad=Alta)
THEN
 JugarTenis=No





5.2. La regla de post-poda (3)

¿Por qué transformar árboles en reglas?

- (a) permite distinguir entre diferentes contextos en el árbol.
- (b) elimina la distinción de los atributos atendiendo a su cercanía a la raíz.
- (c) los árboles frondosos son difíciles de interpretar por los seres humanos.
Las reglas se prefieren en estos casos.

5.2. Atributos con valores continuos

Problema:

ID3 no puede trabajar con valores continuos

Solución:

Discretizar el dominio continuo

5.2. Atributos con valores continuos (2)

Temperatura	6	10	22	30	34	40
JugarTenis	No	No	Si	Si	Si	No

¿Qué valor utilizar para discretizar?

-> [10,22] y [34,40]

seleccionar el que mayor ganancia aporta

Otras alternativas:

- * usar varias particiones (Fayyad e Irani)
- * otras (Utgoff y Brodley; Murthy et al.)

Determinación del umbral

- Se ordenan los valores del atributo continuo de menor a mayor $\{v_1, v_2, \dots, v_m\}$
- Se hace $\forall i h = v_i$, de forma que se parten los valores en 2: $\{v_1, v_2, \dots, v_i\}$ y $\{v_{i+1}, v_{i+2}, \dots, v_m\}$
- Hay por lo tanto $m-1$ posibles umbrales
- Para cada uno de ellos se calcula el valor de la ganancia proporcional
- Se elige aquel umbral con mayor ganancia

UNA VARIANTE DEL EJEMPLO ANTERIOR

Cielo	Temperatura	Humedad	Viento	Clase
Soleado	75	70	Verdadero	Jugar
Soleado	80	90	Verdadero	No Jugar
Soleado	85	85	Falso	No Jugar
Soleado	72	95	Falso	No Jugar
Soleado	69	70	Falso	Jugar
Nublado	72	90	Verdadero	Jugar
Nublado	83	78	Falso	Jugar
Nublado	64	65	Verdadero	Jugar
Nublado	81	75	Falso	Jugar
LLuvia	71	80	Verdadero	No Jugar
LLuvia	65	70	Verdadero	No Jugar
LLuvia	75	80	Falso	Jugar
LLuvia	68	80	Falso	Jugar
LLuvia	70	96	Falso	Jugar

árbol del ejemplo

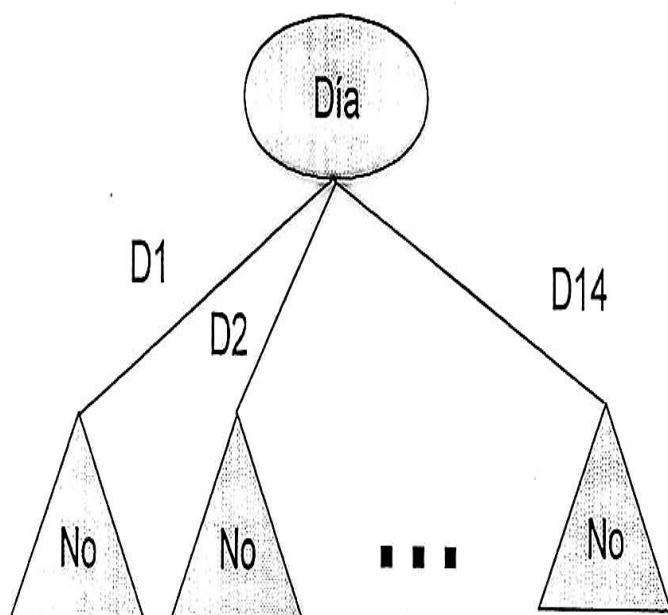
El árbol resultante es:

- Cielo=Soleado
 - Humedad ≤ 75 : Jugar
 - Humedad > 75 : No Jugar
- Cielo=Nublado: Jugar
- Cielo=LLuvioso
 - Viento=Verdadero: No Jugar
 - Viento=Falso: Jugar

▼

5.3. Otras medidas para la selección de atributos

¿Qué ocurre si *Día* se considera como atributo para la construcción del árbol?



5.3. Otras medidas para la selección de atributos (2)

Alternativa: (Quinlan 1986)

ratio de ganancia

$$RG(E,X) = GANANCIA(E,X)/dInfo(E,X)$$

información sobre la división:

$$dInfo(E,X) = -\sum (|E_i|/|E|) \log (E_i/E)$$

Otras alternativas:

Breiman et al.

Dietterich et al.

5.4. Tratamiento de ejemplos con valores perdidos

Problema:

algunos ejemplos contienen valores perdidos.

Solución:

estimar el valor perdido

Alternativas:

- (a) El valor más frecuente (todos los ejemplos)
- (b) El valor más frecuente (subconjunto afect.)
- (c) Asignar probabilidades a cada opción.

METODOLOGÍA CART

El problema de la regresión de funciones ha sido tratado mediante técnicas jerárquicas de división del dominio de entrada.

Algoritmos como CART (Classification And Regression Tree) amplían los métodos para la inducción de árboles de clasificación y decisión (ID3), generalizando el proceso de construcción de clasificadores.

Estas técnicas pretenden construir **una estimación f' de una función desconocida**

$$f: X^R \rightarrow X^Y \quad X^I = X_1 \times X_2 \times \dots \times X_r,$$

partiendo de un conjunto de entrenamiento

$$L = \{(x_i, y_i) \mid i=1 \dots N\} \text{ tal que } f(x_i) = y_i.$$

Suponiendo la continuidad de f en entornos de los valores de entrada x , los modelos que definen a la función estimadora f' se componen de un conjunto de regiones del dominio de entrada, generalmente disjuntas, a las cuales se les asocia un valor de salida:

$$f' = \{(R_i, y_i), R_i \subseteq X, y_i \in Y, i=1,2,\dots,m\}$$

Definido el modelo f' el proceso de inferir un valor de la función $f(x)$ para un x del dominio de entrada consiste en determinar a qué región R_i pertenece x y asignar el valor de salida de dicha región y_i a $f(x)$.

$$f(x) = y_i \text{ si } x \text{ pertenece a } R_i$$

En términos muy generales la construcción jerárquica de estos modelos se establece por medio de una sucesión de modelos f_0' , ..., f_H' construidos por el siguiente algoritmo

- 1.- El modelo f_0' consta de una única región (todo el dominio de entrada X) y una valor de salida asociado que hace mínimo el error cometido entre éste y los datos conocidos L.**
- 2.- Seleccionar del modelo actual, aquella región donde el error cometido sea mayor.**
- 3.- Eliminar la región del modelo.**
- 4.- Dividir la región eliminada en un conjunto de nuevas subregiones y asignar a cada nueva subregión el valor de salida que minimice el error cometido con respecto a los datos conocidos.**
- 5.- Construir un nuevo modelo añadiendo al actual el conjunto de subregiones obtenidas, así como los valores de salida asociados.**
- 6.- Si el nuevo modelo cumple las expectativas sobre el error deseado para la estimación terminar; en caso contrario considerar el nuevo modelo como actual e ir al paso 2.**

La evolución del algoritmo anterior se puede plasmar mediante una estructura arbórea. En este árbol (árbol de clasificación y regresión) los nodos hoja representa cada una de las regiones del modelo actual, mientras que los nodos interiores son aquellas regiones eliminadas de los modelos anteriores.

Las piedras angulares sobre las que descansan el algoritmo anterior, así como los modelos obtenidos son:

1.- La forma de identificar cada región. Las regiones suelen estar definidas como un conjunto de restricciones sobre los valores del dominio de entrada. Estas restricciones se expresan como posibles valores que pueden tomar las variables de entrada o en términos más generales como predicados que se han de verificar.

2.- Valoración del error cometido en la estimación a partir del conjunto de ejemplos L que se dispone. Fijada una función distancia d en el dominio de la salida, varias son las alternativas utilizadas para cuantificar la bondad de la estimación.

Fijadas ambos criterios, el proceso de identificación de la región a dividir y el mecanismo de división de la misma queda totalmente establecido. En general una región quedará dividida al sustituir un predicado sobre alguna de las dimensiones del dominio de entrada que la define, por otro conjunto de nuevos predicados sobre la misma dimensión.

Destacaremos el uso de predicados imprecisos o borrosos en la definición de las regiones del modelo, originando particiones difusas del dominio de entrada. Este tipo de modelos que podríamos denominar como modelos difusos han sido tratados por varios autores estableciendo el concepto de árboles difusos de regresión y decisión.

Con esta idea nosotros hemos desarrollado la mitología denominada ADRI (Árboles Difusos de Regresión e Identificación).

En nuestros modelos se realiza además una selección de características, descubriendo mediante un análisis del error cometido cuales son las variables relevantes para el problema.

Además de la representación mediante árboles de estos modelos, existe la posibilidad de obtener de forma natural otra representación de los mismos mediante sistemas de reglas. El sistema de reglas consistente en una colección de reglas del tipo “ Si antecedente Entonces consecuente ”, donde el antecedente refleja el conjunto de predicados que conforma una región del modelo y el consecuente el valor de salida asociado a dicha región, refleja en su totalidad el modelo considerado.

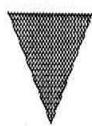
El uso combinado de la representación del modelo mediante sistema de reglas, así como la utilización de la lógica difusa y las variables lingüísticas configuran modelos altamente cualitativos y expresivos que solucionan el problema de la regresión de funciones.

Estas técnicas han sido utilizadas con éxito para la obtención de modelos de funciones no lineales, series caóticas y macroeconómicas.

DISCUSION

- El aprendizaje inductivo con árboles proporciona un método práctico para el aprendizaje de conceptos
- El sesgo inductivo de ID3 tiende a la obtención de árboles pequeños.
- El sobre-ajuste de los ejemplos del conjunto de entrenamiento es un aspecto importante en aprendizaje con árboles de decisión.
- Una gran variedad de extensiones del algoritmo ID3 se han desarrollado.

Nosotros hemos trabajado con árboles de Decisión combinándolos con Lógica Difusa para obtener Reglas Difusas (ADRI).



Bibliografía

Quinlan, J.R. C4.5: Programs for Machine Learning, 1993.

Mitchell, T. Machine Learning, McGraw-Hill, 1997. (Capítulo 3)

Russell, S. y Norvig, P. Inteligencia Artificial: un enfoque moderno, Prentice Hall, 1996. (Capítulo 18)

Bibliografía complementaria

- Breiman, L., Friedman, J.H., Olshen, R.A., Stone, P.J., *Classification and regression trees*. Belmont, CA: Wadsworth International Group. (1984).
- Dietterich, T.G., Kearns, M., y Mansour, Y., *Applying the weak learning framework to understand and improve C4.5*. Proceedings of the 13th International Conference on Machine Learning, pp. 96-104. San Francisco: Morgan Kaufmann. (1996).
- Fayyad, U.M, Weir, N., y Djorgovski, S., *Multi-interval discretization of continuous-valued attributes for classification learning*. In R. Bajcsy (Ed.), *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, (pp. 1022-1027). Morgan-Kaufmann. (1993).
- Murphy, P.M., Kasif, S., Salzberg, S., *A system for induction of oblique decision trees*, *Journal of Artificial Intelligence Research*, 1, pp. 257-275. (1994).
- Quinlan, J.R., *Induction of decision trees*. *Machine Learning*, 1(1), pp. 81-106. (1986).
- Quinlan, J.R., *Rule induction with statistical data--a comparison with multiple regression*. *Journal of the Operational Research Society*, 38, pp. 347-352. (1987)
- Utgoff, P.E, y Brodley, C.E., *Incremental induction of decision trees*, *Machine Learning*, 4(2), pp. 161-186. (1991).