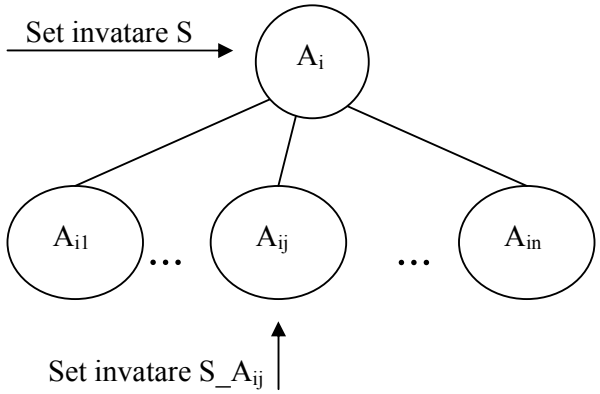
Algoritmul ID3

Algoritmul ID3 [1], [2] este un algoritm de clasificare supervizata, ce construieste un arbore de decizie.

Multimea de invatare S este formata dintr-un set de obiecte, fiecare obiect fiind caracterizat printr-o multime de atribute (A). Fiecare obiect face parte dintr-o clasa C. Scopul algoritmului este de a construi un arbore de decizie care sa poata clasifica orice obiect (specificat prin acelasi set de atribute A).

Constructia unui nod in arborele de decizie (pentru cazul in care atributele au asociate numai valori discrete):



Se partitioneaza setul de invatare S in subseturile S_A_{i1}, S_A_{i2}, ..., S_A_{in}, corespunzator numarului de valori ale atributului A_i.

Un nod din arbore va fi asociat atributului A_i, pentru care $Gain(A_i)$ are valoare maxima (in raport cu toate atributele). Nodul corespunzator atributului A_i va avea n fii, unde n reprezinta numarul de valori posibile ale atributului A_i.

$$Gain(A_i) = Entropy(S) - \sum_{j=1}^{n} \frac{|S\_A_{ij}|}{|S|} \; Entropy(S\_A_{ij})$$

Entropia pentru 2 clase *p* si *n*:

$$Entropy(S) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Calcul entropie (caz general):

$$Entropy\,(S) = \Sigma\text{-}P(Clasa=C_i)*\log_2(P(Clasa=C_i)$$

Exemplu:

Se considera setul de invatare S, in care fiecare obiect este caracterizat prin atributele age, competition, type, care au valorile posibile:

ATRIBUT	VALORI POSIBILE
age	old, midlife, new
competition	no, yes
type	swr, hwr

Setul S de invatare este:

AGE	COMPETITION	TYPE	PROFIT
old	yes	swr	down
old	no	swr	down
old	no	hwr	down
mid	yes	swr	down
mid	yes	hwr	down
mid	no	hwr	up
mid	no	swr	up
new	yes	swr	up
new	no	hwr	up
new	no	swr	up

$$\text{Entropy(S)} = -\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10} = 1$$

Split Age: ==>3 fii:

- R1 (Age = new) (0 down, 3 up)
- R2 (Age = mid) (2 down, 2 up)
- R3 (Age = old) (3 down, 0 up)

$$Entropy(S\_R1) = 0$$

$$Entropy(S\_R2) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$Entropy(S\_R3) = 0$$

$$Gain(Age) = 1 - (0 + \frac{4}{10} + 0) = 0.6$$

Split Type: ==>2 fii:

- R1 (Type =swr) (3 down, 3 up)
- R2 (Type =hwr) (2 down, 2 up)

$$Entropy(S\_R1) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6} = 1$$

$$Entropy(S\_R2) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4} = 1$$

$$Gain(Type) = 1 - (\frac{6}{10} + \frac{4}{10}) = 0$$

Split Competition: ==>2 fii:

- R1 (Competition = yes) (3 down, 1 up)
- R2 (Competition = no) (2 down, 4up)

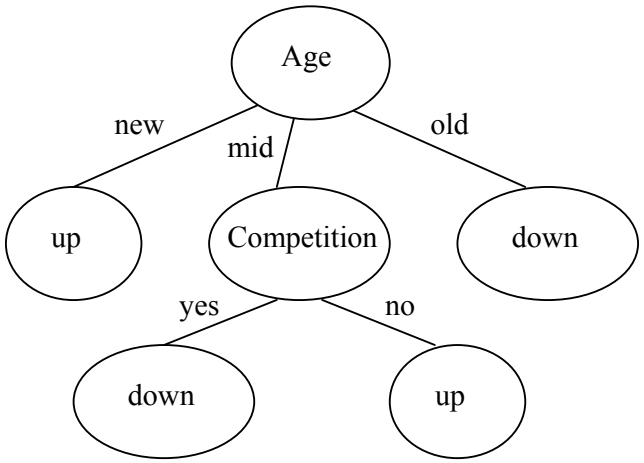
$$Entropy(S\_R1) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4} = 0.81$$

$$Entropy(S\_R2) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6} = 0.91$$

$$Gain(Competition) = 1 - (\frac{4}{10} * 0.81 \quad + \frac{6}{10} * 0.91) = 0.13$$

Max(Gain(Age), Gain(Type), Gain(Competition)) = Gain(Age) ==> atributul corespunzator radacinii este Age.

Arborele de decizie construit este:



Alt exemplu de constructie arbore de decizie: [3]

Resurse

[1] http://en.wikipedia.org/wiki/ID3_algorithm
[2] http://cs.nyu.edu/faculty/davise/ai/id3.pdf
[3] http://cs.nyu.edu/faculty/davise/ai/id3-ex.txt