ARBORI DE DECIZIE

ALTE APLICATII IN CARE SE POATE FOLOSI CLASIFICAREA

- Clasificarea clientilor unei banci care aplica pentru credit bancar in “fara riscuri” sau “cu riscuri”
- Clasificarea clientilor unui magazin de produse electronice in posibili cumparatori de calculatoare sau nu

ARBORII DE DECIZIE

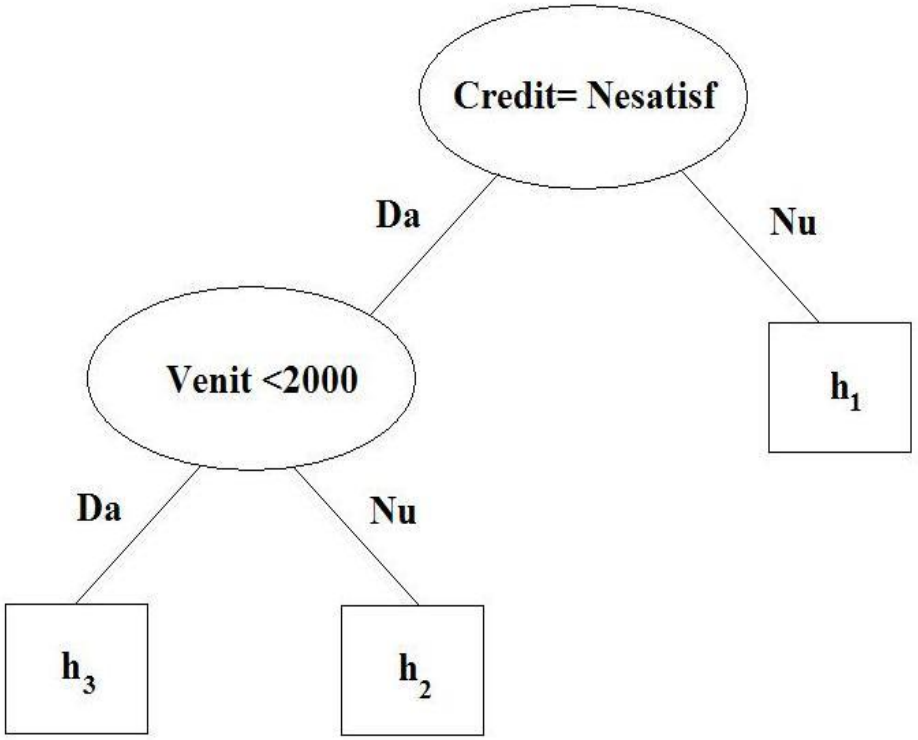
- A***rbore de decizie*** este o structura care poate fi folosita pentru a imparti o colectie mare si eterogena de exemple intr-un sir de colectii din ce in ce mai mici si mai omogene in raport cu un atribut tinta.
- Divizarea colectiei se face prin aplicarea unui sir de reguli de decizie simple.
- Ca orice arbore in teoria grafurilor, arborele de decizie are drept componente
  - noduri,
  - ramuri,
  - frunze

si se reprezinta cu ramurile in jos, plecand de la radacina.

- Nodurile interne= teste facute colectiei de date in functie de valorile unui atribut
- Ramurile = valori posibile ale testelor
- Frunzele = modurile de clasificare (clasa careia ii apartine colectia de inregistrari din nodul respectiv)

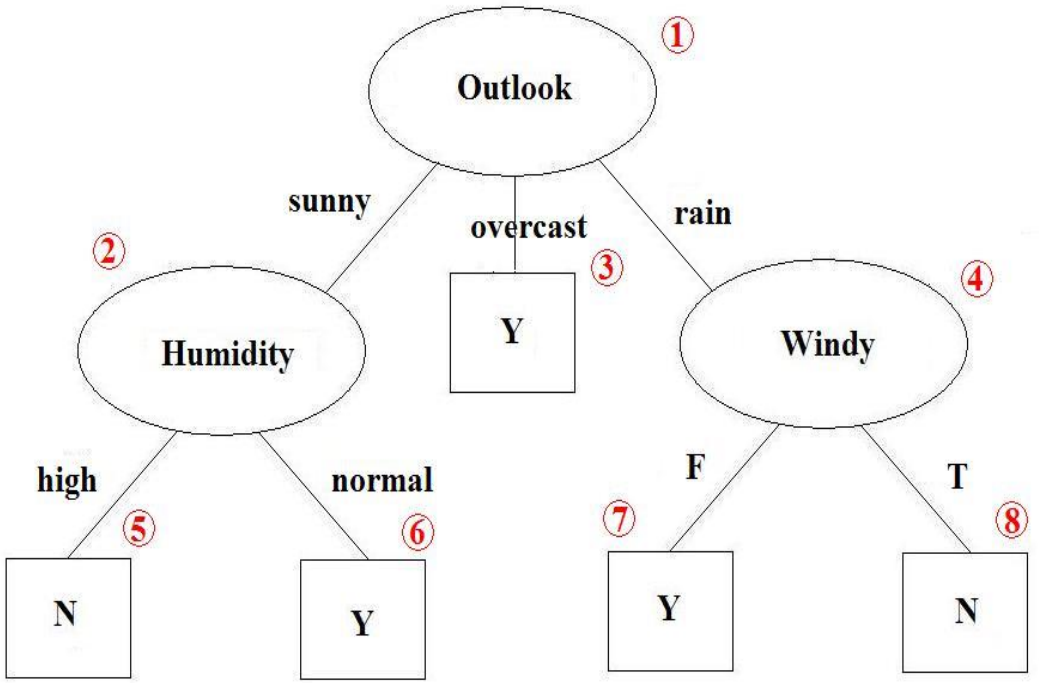
EXEMPLU 1

- Pentru baza de date pentru determinarea acordarii unui credit, un arbore de decizie asociat este:



EXEMPLU 2

- Un arbore de decizie pentru baza de date corespunzatoare meciului de baseball in care atributul tinta este “Play”, este:



- In procesul decizional, o inregistrare intra in arbore la radacina si iese printr-un nod terminal clasificata.
- Un drum de la radacina la o frunza este o expresie care exprima regula folosita pt clasificarea inregistrarii.
- De exemplu, inregistrarea: {sunny, cool, high, T} parcurge in ordine nodurile 1, 2, 5 si iese cu clasificarea Play=N, regula folosita fiind: {Outlook= sunny si Humidity=high}.
- Frunze diferite pot fi clasificate la fel, chiar daca din motive diferite. De exemplu, nodurile 3 si 6 sunt clasificate la fel.
- Daca atributul tinta este o var discreta atunci arb de decizie se numeste arbore de clasificare.

CONSTRUIREA ARBORELUI DE DECIZIE

- Arborele este construit de sus in jos recursiv, in maniera Divide et Impera. Atributele de intrare sunt discretizate in prealabil.
- La inceput, in radacina arborelui se afla toate inreg multimii de training.
- Se selecteaza atributul care da cea mai buna impartire a nodului radacina.
- Se partitioneaza multimea datelor conform valorilor testului efectuat asupra atributului selectat.
- Pt. fiecare partitie se repeta pasii de mai sus.
- Conditii de oprire a divizarii unui nod:
  - Toate inreg nodului apartin aceleasi clase
  - Nu mai sunt atribute pt a putea face divizarea (se alege clasa cu cele mai multe inregistrari)
  - Nu mai exista inregistrari.

CRITERII PENTRU ALEGEREA CELUI MAI BUN ATRIBUT

- Cel mai bun atribut= atributul care da cea mai buna divizare a unui nod
- Criterii – arbore de clasificare
  - Indicele Gini
  - Entropia (sau informatia dobandita)
  - Testul Chi-patrat

INDICELE GINI

- Daca T este o multime de date ce contine inreg din n clase (deci atributul tinta are n valori discrete posibile), indicele Gini se defineste:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

unde p_j = frecventa relative a clasei j in T

$$p_j = \frac{nr\ aparitii\ ale\ clasei\ j}{nr\ inreg\ ale\ colectiei\ T}$$

- Daca T se imparte in submultimile astfel incat *T* are *N* elemente, T_1 are N_1 elemente, …, T_k are N_k elemente, atunci

  indicele *Gini* al diviziunii

$$gini_{diviziune}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2) + \ldots \frac{N_k}{N} gini(T_k)$$

- Indicele Gini ne da probabilitatea ca doua elemente alese la intamplare din colectia data sa nu fie in aceeasi clasa.
- Observati ca daca T contine numai inreg dintr-o clasa atunci

$$gini(T) = 1 - \left[\left(\frac{0}{N} \right)^2 + \ldots + \left(\frac{0}{N} \right)^2 + \left(\frac{N}{N} \right)^2 \right] = 0$$

Daca n =2 si T contine acelasi numar de inreg din clasa 1 ca si din clasa 2 atunci

$$gini(T) = 1 - \left[\left(\frac{N/2}{N} \right)^2 + \left(\frac{N/2}{N} \right)^2 \right] = 1 - 1/2 = 1/2$$

CEL MAI BUN ATRIBUT

- Cel mai bun atribut = cel pentru care indice $gini_{diviziune}(T)$ este **cel mai mic.**

APLICATIE

- Construirea arborelui de decizie corespunzator bazei de date asociate datelor meteorologice ale unor zile. Atributul tinta este “Play” cu valorile Y/N și corespunde deciziei dacă într-o zi dată sunt condiții favorabile unui joc de baseball.

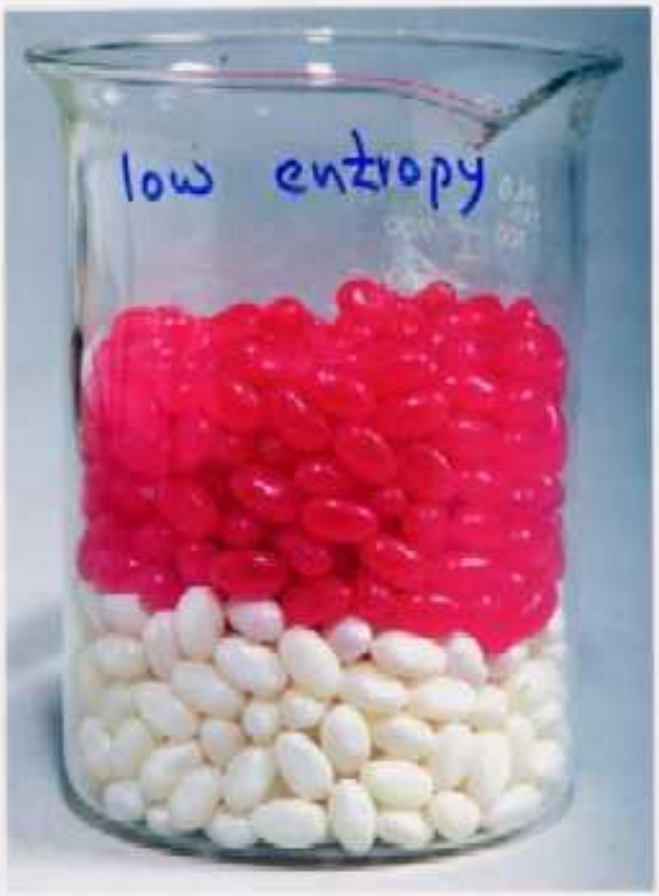
MULTIMEA DE TRAINING

Outlook	Temp	Humidity	Windy	Play
sunny	hot	high	F	N
sunny	hot	high	T	N
overcast	hot	high	F	Y
rain	mild	high	F	Y
rain	cool	normal	F	Y
rain	cool	normal	T	N
overcast	cool	normal	T	Y
sunny	mild	high	F	N
sunny	cool	normal	F	Y
rain	mild	normal	F	Y
sunny	mild	normal	T	Y
overcast	mild	high	T	Y
overcast	hot	normal	F	Y
rain	mild	high	T	N

ENTROPIA (SAU INFORMATIA DOBANDITA)

- Entropia este o masura a modului de dezorganizare a unui sistem.

ENTROPY IS AN INDICATOR OF RANDOMNESS



8.2a

8.2b

- Observati ca daca T contine numai inreg dintr-o clasa atunci

$$entropie(T) = -1\log(1) = 0$$

Daca n =2 si T contine acelasi numar de inreg din clasa 1 ca si din clasa 2 atunci

$$entropie(T) = -1/2\log(1/2) - 1/2\log(1/2) = -\log(1/2) = 1$$

ENTROPIA (SAU INFORMATIA DOBANDITA)

- Daca T este o multime de date ce contine inreg din n clase (deci atributul tinta are n valori discrete posibile),entropia se defineste:

$$\text{entropie}(T) = -p_1 \log_2 p_1 - p_2 \log_2 p_2 \ldots - p_n \log_2 p_n$$

unde pj = frecventa relative a clasei j in T,

$$p_j = \frac{nr\ aparitii\ ale\ clasei\ \ j}{nr\ inreg\ \ ale\ colectiei\ T}$$

- Daca T se imparte in submultimile astfel incat *T* are *N* elemente, T_1 are N_1 elemente, …, T_k are N_k elemente, atunci

$$entropie_{diviziune}(T) = \frac{N_1}{N} entropie(T_1) + \frac{N_2}{N} entropie(T_2) + ... \frac{N_k}{N} entropie(T_k)$$

- Definim informatia dobandita a diviziunii:

$$INFO_{divizare} = entropie(T) - entropie_{divizare}(T) = entropie(T) - \sum_{j=1}^{k} \frac{N_j}{N} entropie(T_j)$$

CEL MAI BUN ATRIBUT

- Cel mai bun atribut = cel pentru care $INFO_{diviziune}(T)$ este **cel mai mare.**