

TEHNICI DE EXPLORARE A DATELOR -DATA MINING -

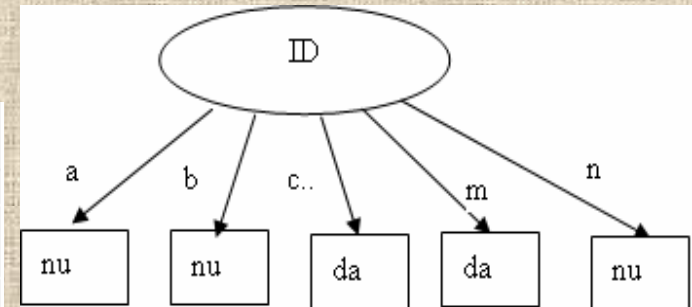
CURS 6

Atribute cu un număr mare de valori distincte

- Pot genera un număr mare de noduri copil
=> apar probleme cu calculul câștigului de informație
- Exemplu – cazul extrem cand un atribut are câte o valoare diferită pentru fiecare instanță din mulțimea de date

Exemplu

ID	Aspect	Temperatura	Umiditate	Vant	Desfasurare joc
a	insorit	cald	mare	F	Nu
b	insorit	cald	mare	A	Nu
c	innorat	cald	mare	F	Da
d	ploios	medie	mare	F	Da
e	ploios	scazuta	normala	F	Da
f	ploios	scazuta	normala	A	Nu
g	innorat	scazuta	normala	A	Da
h	insorit	medie	mare	F	Nu
i	insorit	scazuta	normala	F	Da
j	ploios	medie	normala	F	Da
k	insorit	medie	normala	A	Da
l	innorat	medie	mare	A	Da
m	innorat	cald	normala	F	Da
n	ploios	medie	mare	T	Nu



- Informația solicitată pt a specifica clasa data de o valoare a acestui atribut este:

$$\text{Info}([0,1]) + \text{Info}([0,1]) + \text{Info}([1,0]) + \dots + \text{Info}([1,0]) + \text{Info}([0,1]) = 0$$

- Câștigul de informație pentru acest atribut este chiar informația din rădăcină **$\text{Info}([9,5])=0,940$ biti** – este mai mare decât câștigul calculat pentru orice alt atribut => ID ar fi inevitabil ales ca atribut de test
 - ramificarea pe ID nu ajuta la clasificarea unor cazuri noi si nu dă nici o indicație referitoare la structura deciziei

- SOLUTIA: este necesara o noua masura – ***rata castigului (gain ratio)***
 - ia in considerare numarul si dimensiunea nodurilor copil in care un atribut de test divizeaza setul de date, fara a considera informatia referitoare la clasa
- In exemplul considerat valoarea ***informatiei pentru divizare*** este:
 $\text{Info}([1,1,1,...,1,1]) = -1/14 * \log 1/14 * 14$ (aceeasi fractie – 1/14) apare de 14 ori)
 $\text{Info}([1,1,1,...,1,1]) = \log 14 = 3,807$ biti (valoare foarte mare)
- Rata castigului se calculeaza prin impartirea castigului de informatie (0,940) la valoarea informatiei de divizare pentru atribut (3,807)
- => pentru atributul ID :
 $\text{gain ratio} = 0,940 / 3,807 = 0,247$

Daca se calculeaza aceasta rata pentru atributul *aspect* , care divizeaza setul de date in seturi de dimensiune 5,4,5 avem:

Informatia de divizare (***numarul de biti necesari pentru a determinarea ramurii i se asigneaza fiecare instanta***)

$$\text{Info}([5,4,5]) = 1,577$$

$$\text{Gain ratio} = \text{info gain} / \text{split info} = (0,94 - 0,693) / 1,577 = 0,157$$

	Aspect	Temperatura	Uniditate	Vant
Info	0,693	0,911	0,788	0,892
Info gain	0,247	0,029	0,152	0,048
Split info	1,577	1,557	1,000	0,985
Gain ratio	0,157	0,019	0,152	0,049

Implementarea arborilor de decizie in cazul masinilor reale

- ID3 algoritm de baza pentru constructia arborilor de decizie
- C4.5 si C 5.0 completeaza aspectele intalnite in realitate care nu sunt acoperite prin cerintele ID3
 - Atribute numerice (ID3 solicita valori descriptive)
 - Valori lipsa
 - Metode de taiere a arborilor

Atribute cu valori numerice

- Cea mai simpla metoda este cea prin care se restrange posibilitatile de diviziune a unui nod, in doua parti (diviziune binara)
 - => diferite intre modul in care sunt tratate attributele numerice si cele nominale

Atribute nominale	Atribute numerice
1. Determina divizarea nodului in atatea ramuri cate valori distincte are atributul	1. Se foloseste divizarea binara
2. Se foloseste intreaga informatie oferita intr-un singur nod	2. Divizarile succesive pot continua sa furnizeze informatii in orice nod in care se face divizarea
3. Poate fi testat o singura data pe orice cale de la radacina la frunza	3. Poate fi testat de mai multe ori pe o cale => genereaza arbori greu de interpretat

Exemplu

- Se modifica setul de date referitor la vreme astfel incat attributele *temperatura* si *umiditate* sa aiba valori numerice.

Aspect	Temperatura	Umiditate	Vant	Desfasurare joc
insonit	85	85	F	Nu
insonit	80	90	A	Nu
innorat	83	86	F	Da
ploios	70	96	F	Da
ploios	68	80	F	Da
ploios	65	70	A	Nu
innorat	64	65	A	Da
insonit	72	95	F	Nu
insonit	69	70	F	Da
ploios	75	80	F	Da
insonit	75	70	A	Da
innorat	72	90	A	Da
innorat	81	75	F	Da
ploios	71	91	A	Nu

- Daca se considera temperatura ca atribut de divizare atunci valorile implicate si clasele corespunzatoare sunt:

64	65	68	69	70	71	72	75	80	81	83	85
da	nu	da	da	da	nu	nu	da	nu	da	da	nu
						da	da				

- Exista cel mult 11 posibilitati de a diviza intervalul considerat de temperaturi
 - Daca clasele identice sunt considerate ca apartinand unui interval atunci gama de temperaturi se poate diviza in 8 moduri
- Castigul de informatie pentru fiecare poate fi calculat dupa formula cunoscuta

Alternative

- Sunt mai greu de realizat
- Produc arbori mai usor de inteles
- Sa fie permisa o testare mai complexa a atributelor numerice prin compararea valorilor acestora cu mai multe constante intr-un singur nod al arborelui
- Atributul sa fie supus unei operatii de pregatire prin discretizare

Valori lipsa

- Sunt foarte des intalnite in multimile de valori reale
- Pot fi tratate ca o alta valoare posibila pentru atributul considerat
 - aplicabil daca lipsa atributului afecteaza semnificativ rezultatele
 - Daca nu exista o semnificatie anume a faptului ca pentru o anumita instanta lipseste valoarea atributului-> este necesara o solutie mai subtila
 - Este tentant sa se ignore pur si simplu toate instantele in care valori ale atributelor lipsesc
 - De multe ori solutia este neviabila => instantele cu valori lipsa pot oferi informatii pretioase
- Daca attributele cu valori lipsa nu iau parte la decizii aceste instante sunt tratate ca oricare altele

Valori lipsa

- Problema 1:
 - Cum se aplica un arbore de decizie dat unei instante in care anumite attribute testate au valori lipsa?
- *Posibila solutie*: impartirea speculativa a instantei in bucati utilizand o metoda de ponderare numerica si trimiterea catre frunzele arborelui a unor parti proportionale cu numarul instantelor de antrenare care merg pe ramura respectiva
 - Eventual, diferite parti ale instantei vor ajunge, fiecare intr-un nod frunza si deciziile din aceste noduri frunza vor trebui recombinate folosind ponderile care au filtrat aceste frunze
- Se poate aplica acelasi mod de calcul pentru castigul de informatie si pentru rata castigului, instantelor partiale
 - Si in acest caz se vor folosi ponderile

Valori lipsa

- Problema 2:
 - Cum trebuie partitionata multimea e antrenare odata ce a fost ales un atribut de divizare, pentru a permite aplicarea recursiva a procedurii de inductie a arborelui de decizie in fiecare nod copil?
- *Posibila solutie*: se foloseste aceeaasi procedura de ponderare ca cea prezentata anterior
 - Datele pot fi divizate in noduri pe niveluri inferioare chiar daca valorile altor attribute nu sunt in totalitate cunoscute
 - Partile instantei contribuie la decizii la nivelul nodurilor inferioare prin calculul castigului de informatie care, de aceasta data va fi ponderat

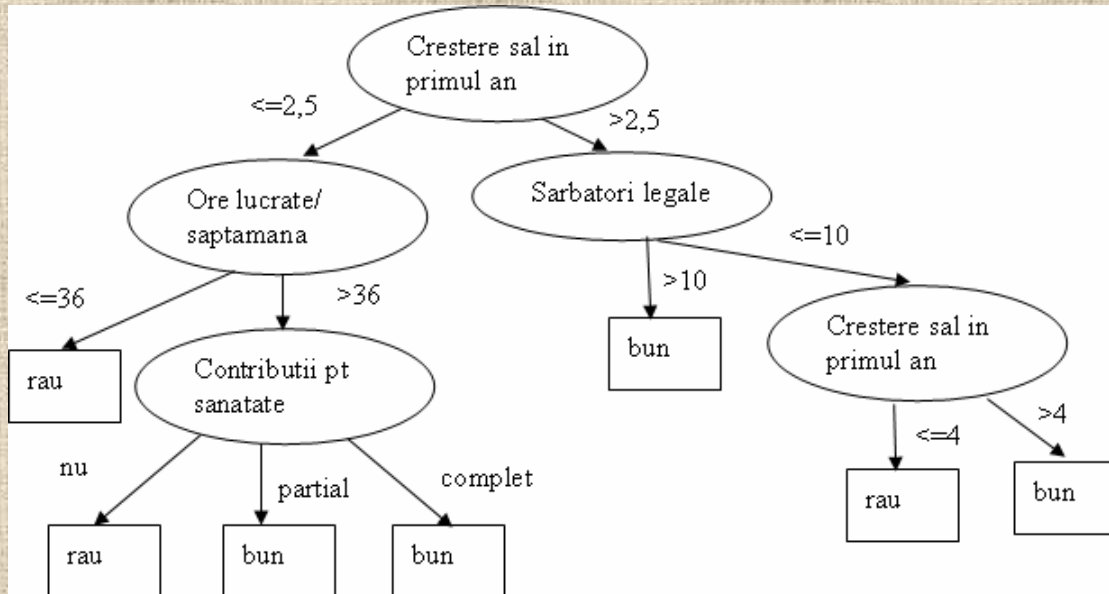
Taierea arborilor

- De multe ori un arbore de decizie mai simplu ofera solutii mai bune decat unu complex
- Taierea (pruning) unui arbore il aduce la o forma mai simpla care sa permita o clasificare optima
- Exista doua strategii posibile de taiere
 - **Postpruning** (backward pruning)
 - Presupune inductia completa a arborelui urmata de taierea subarborilor fara relevanta
 - Avantaj –sunt situatii in care doua attribute, considerate individual par sa nu contribuie la clasificare, dar sunt un predictor puernic atunci cand sunt luate impreuna
 - **Prepruning** (forward pruning)
 - Implica incercarea de a decide pe durata procesului de constructie a arborelui momentul de la care nu mai este necesara dezvoltarea de subarbori
 - Se evita construirea subarborilor care vor fi ulterior taiati

Taierea arborilor- postpruning

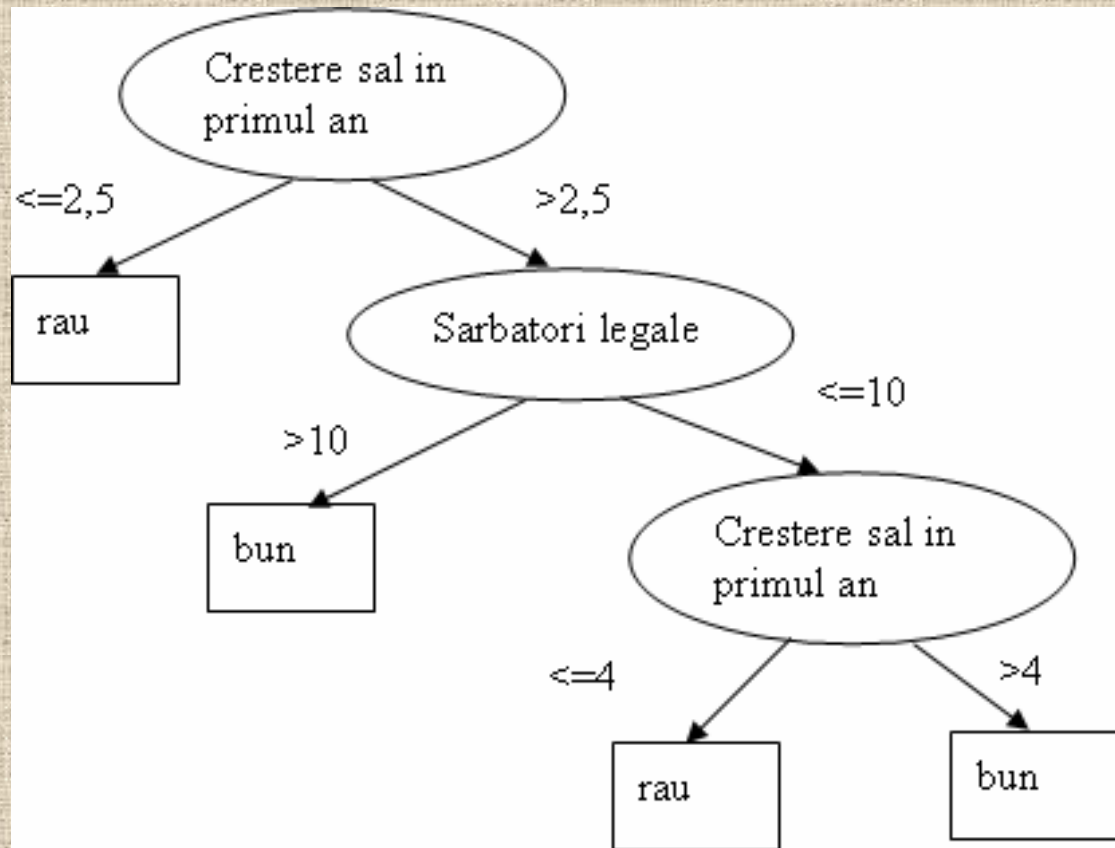
- Sunt posibile 2 operatii:
 - Inlocuirea subarborilor cu frunze
 - Este principala operatie de taiere
 - Poate determina o descrestere a acuratetii pe setul de antrenare, daca arborele original a fost produs de algoritmul de constructie a arborilor de decizie, deoarece acesta putea continua constructia arborelui pana in momentul in care toate nodurile devin pure sau pana la testarea tuturor atributelor
 - Operatia incepe de la frunze si se continua catre radacina arborelui
 - De regula se face in mai multe etape
 - Cresterea subarborilor
 - Este o operatie mai complexa si mai costisitoare
 - Folosita in decizii de constructie a arborilor in sistemul C4.5.
 - Este restrictionata la cresterea subarborelui celei mai populare ramuri
- In fiecare nod, o schema de invatare ar trebui sa decida cand ar trebui facuta o inlocuire de subarbore, o crestere de subarbore sau cand, subarboarele ar trebui lasat asa cum este

Taierea arborilor – postpruning- inlocuirea subarborilor - exemplu



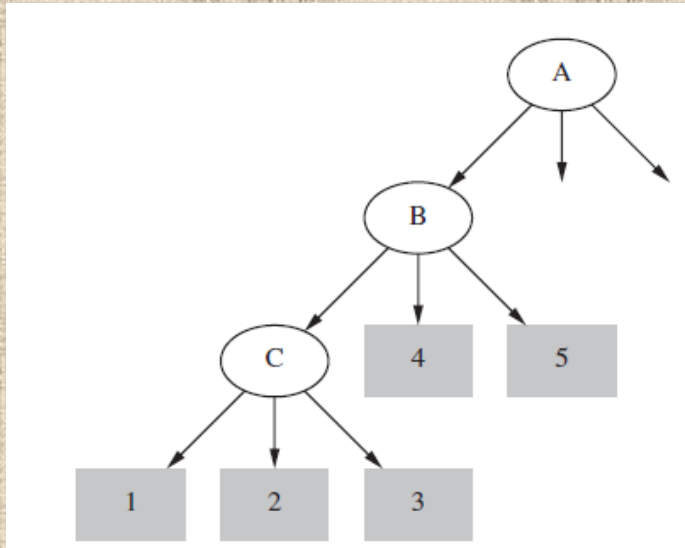
- Se ia mai intai in calcul inlocuirea celor trei noduri copil din subarborele “contributii pentru sanatate” cu o singura frunza
- Apoi, continuand lucrul dinspre frunze ar trebui considerata inlocuirea subarborelui “ore lucrate/saptamana”, care are acum 2 noduri copil, cu un singur nod frunza
- In final, s-ar putea considera inlocuirea celor 2 noduri copil din subarborele “crestere sal in primul an” cu un singur nod frunza
 - Aceasta decizie nu a fost luata => arborele a ramas la stadiul anterior si arata.....

Taierea arborilor – postpruning- inlocuirea subarborilor - exemplu



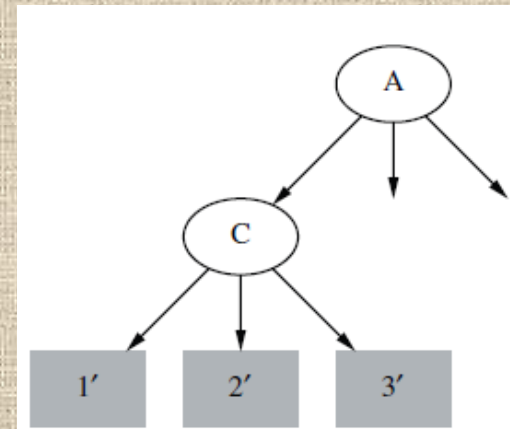
- Intregul subarbore care implica 2 noduri interne si 4 noduri frunza a fost inlocuit printr-o singura frunza "rau"

Taierea arborilor – postpruning- cresterea subarborilor - exemplu



- Intregul subarbore din nodul C va fi “crescut” pentru a inlocui subarboarele din nodul B
- In ac caz copii din nodurile B si C sunt frunze, dar la modul general se pot considera a fi tot subarbori
- Se considera ca ramura de la B la C are mai multe exemple de antrenare decat ramura de la nodul B la nodul 4 sau la nodul 5

- Este necesara reclasificarea exemplelor din nodurile marcate 4 si 5 in noi subarbori care pornesc de la nodul C
 - de aici vine schimbarea notatilor din 1, 2 3 in 1', 2' respectiv 3'



Estimarea ratelor de eroare

- pentru a decide cand este indicata taierea unui arbore si ce metoda este mai potrivita
 - Necesare estimarea ratei de eroare care ar putea fi asteptata intr-un anumit nod pentru un anumit set de date de test independent alese (De ce set de test si nu setul de antrenare?)
 - Estimarea trebuie facuta atat in nodurile interne cat si in cele frunza
 - Poate indica clar cand trebuie inlocuit sau crescut un subabore prin simpla comparare a erorii estimate a arborilor cu cea data de inlocuitorul propus pentru acestia

Estimarea erorii

- Inainte de a estima eroarea pentru un subarbore propus pentru crestere, exemplele din nodurile care dispar ar trebui reclasificate temporar in arborele dezvoltat
- Un mod uzual de a calcula eroarea estimata este tehnica standard de verificare
 - Se pastreaza o parte din datele originale si se utilizeaza ca o multime independenta de test ce este folosita pentru estimarea erorii in fiecare nod
 - Dezavantaj – constructia arborelui se realizeaza pe o multime redusa de date
- alternativa – incercarea de a face estimari de erori chiar pe baza setului de antrenare
 - metoda utilizata in C4.5.
 - este o abordare euristica bazata pe rationamente statistice care functioneaza destulde bine in practica

De la arbori, la reguli

- Este posibila citirea unei reguli direct dintr-un arbore de decizie
 - Se genereaza cate o regula pentru fiecare frunza prin conjunctia tuturor testelor intalnite de-a lungul unei cai de la radacina la frunze
 - => reguli clare pentru care nu are importanta ordinea de executie
 - Uneori acestea sunt mai complexe decat este necesar iar rata estimata de eroare furnizeaza exact mecanismul necesar pentru taierea regulilor

- Pentru o regula data se considera fiecare conditie ca un candidat la eliminare
 - Se elimina conditia vizata
 - Se verifica ce exemple de antrenare sunt acoperite de regula rezultata
 - Se calculeaza pentrua acesata o rata pesimista de eroare care se compara cu estimarea pesimista pentru regula originala
 - Daca noua regula este mai buna se sterge definitiv conditia testata si se cauta o noua conditie ce va si stearsa
 - Procesul se incheie cand nu mai exista nici o conditie candidata la stergere
- Dupa ce toate regulile au fost taiate in acest mod, este necesar sa se verifice daca exista duplicate in setul de reguli rezultat, care vor trebui la randul lor eliminate
- Metoda **costisitoare** de a detecta conditiile redundante, care nici **nu garanteaza** ca se va obtine cel mai potrivit set de reguli

- **Imbunatatiri:**

- S-ar putea lua in considerare, pentru eliminare, toate submultimile de conditii -> mult prea scump
- Utilizarea unei tehnici de optimizare cum ar fi algoritmi genetici pentru a selecta cea mai buna versiune a regulii

- **Problema (in toate cazurile) – costul calculului**

- Pentru fiecare conditie candidata la stergere, efectul regulii trebuie reevaluat pentru toate instantele de antrenare => generarea regulilor din arbore poate fi un proces foarte lent

Metode de constructie a arborilor de decizie

Metodă	Caracteristici
CART	Divizare binară bazată pe GINI (partiționare recursivă motivată printr-o predicție statistică), are două ramuri de la oricare nod neterminal (care nu e nod frunză). Tăierea este bazată pe măsurarea complexității arborelui.. Suportă clasificarea și regresia. Tratează variabile continue, și necesită pregătirea prealabilă a datelor.
ID3/C4.5 și C5.0	Produce arbori cu ramuri multiple pentru un nod. Numărul de ramuri este egal cu numărul de categorii de preziceri. Combină arbori decizionali multipli într-un singur clasificator. Utilizează câștigul de informație pentru divizare. Tăierea este bazată pe o rată de eroare la fiecare frunză.
CHAID	Divizarea utilizează teste χ^2 (detectarea relațiilor statistice complexe). Numărul de ramuri variază de la două la numărul de categorii de preziceri.
SLIQ	Clasficator scalabil rapid. Algoritm rapid de tăiere a arborelui.
SPRINT	Pentru seturi de date mari. Divizarea este bazată pe valoarea unui singur atribut. Înlătură toate restricțiile de memorie prin utilizarea structurilor de date tip listă de attribute.
RAIN FOREST	Construiește o listă de tip AVC (attribute, valori , clase). Separă aspectele de scalabilitate de criteriile care determină calitatea arborelui.

Tabel 6.1. Diferențe ale metodelor de inducție a arborilor de decizie.