

TEHNICI DE EXPLORARE A DATELOR -DATA MINING -

CURS 4

Reprezentarea cunostintelor

- In proc KDD termenul de cunostinta este asociat cu tiparele descoperite in pasul de data mining
- Exista mai multe moduri de reprezentare a tiparelor descoperite prin proc de KDD
 - fiecare din aceste moduri de reprezentare necesita tipuri diferite de tehnici pentru a fi generate

Tabele de decizie

- Cel mai simplu mod de reprezentare a rezultatelor unui proces de invatare

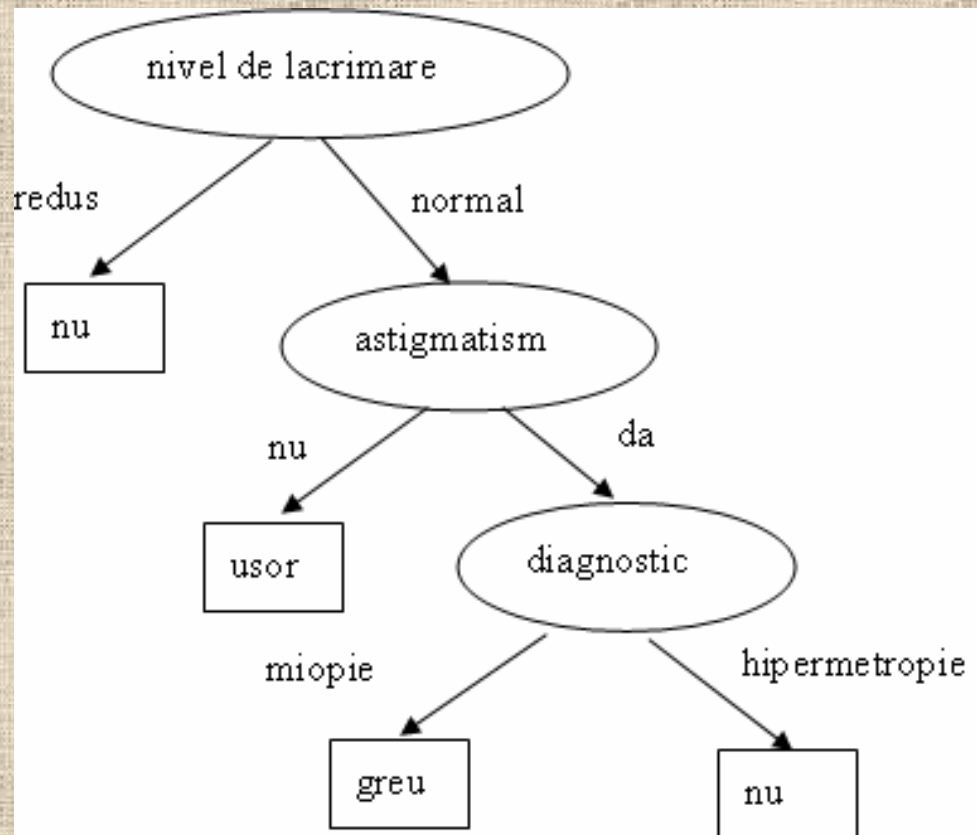
Aspect	Temperatura	Umiditate	Vant	Desfasurare joc
insorit	cald	mare	F	Nu
insorit	cald	mare	A	Nu
innorat	cald	mare	F	Da
ploios	medie	mare	F	Da
ploios	scazuta	normala	F	Da
ploios	scazuta	normala	A	Nu
innorat	scazuta	normala	A	Da
insorit	medie	mare	F	Nu
insorit	scazuta	normala	F	Da
ploios	medie	normala	F	Da
insorit	medie	normala	T	Da
innorat	medie	mare	T	Da
innorat	cald	normala	F	Da
ploios	medie	mare	T	Nu

Problema deciziei de desfasurare a unui joc tinand cont de conditiile meteo

Arbori de decizie

- Solutia naturala de reprezentare a cunostintelor, daca se face o abordare “divide si cucereste” a unei probleme de invatare dintr-un set de instante independente
 - Nodurile unui arbora de decizie implica testarea unor atribute particulare
 - in mod curent se compara valoarea unui atribut cu o constanta
 - pot fi comparate si valorile a doua atribute
 - se pot folosi functii pentru unul sau mai multe atribute
 - nodurile frunza dau:
 - o clasificare care se aplica tuturor instantelor care ajung in frunza
 - o multime de clasificari
 - o distributie de probabilitati pentru toate clasificarile posibile
 - Clasificarea unui nou caz, presupune ca:
 - acesta sa parcurga arborele corespunzator cu valorile atributelor testate in noduri succesive
 - Cand se ajunge intr-un nod frunza instanta este clasificata in acord cu clasa asociata

Arbori de decizie (2)



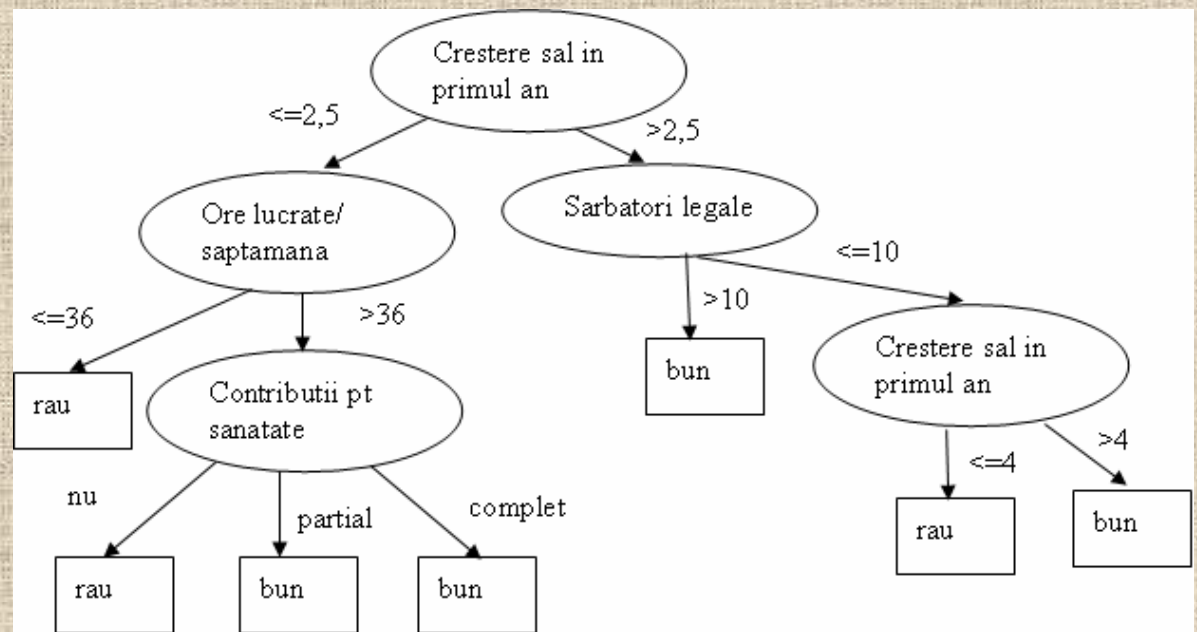
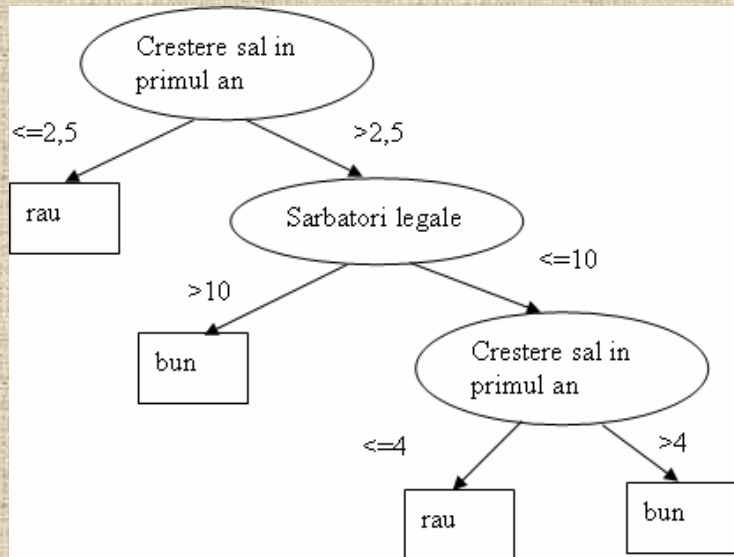
Problema masurii in care sunt suportate lentile de contact

Arbori de decizie (3)

Atribut	UM	1	2	3	40
perioada	ani	1	2	3 ...	2
Crestere de salariu in primul an	procent	2%	4%	4,3%	4,5%
Crestere de salariu in al doilea an	procent	?	5%	4,4%	4%
Crestere de salariu in al treilea an	procent	?	?	?	?
Ore de lucru /saptamana	ore	28	35	38	40
Sporuri de ture	procent	?	5%	4%	4%
Posibilitati de perfectionare	da/nu	da	?	?	?
Sarbatori legale	zile	11	15	12	12
Contributii pentru sanatate	nu/partial/complet	nu	?	complet	partial
Acceptabilitatea contractului	bun/rau	rau	bun	bun	bun

Problema alegerii semnarii unui contract de munca

Arbori de decizie (4)



Arbori de decizie (5)

- Daca atributul este numeric:
 - de regula se verifica daca acesta este $>$ sau $<$ decat o constanta predeterminata
 - \Rightarrow un mod de divizare in doua
 - Exista si posibilitatea de a diviza un nod in trei:
 - Daca valorile lipsa sunt tratate ca o valoare determinata a unui atribut se creează o a treia ramura
 - Se pot testa trei conditii pentru o valoare a unui atribut: daca este $>$, $=$ sau $<$ decat o valoare predeterminata
 - Pentru attributele pentru care opțiunea “egal” nu are semnificatie s-ar putea face compararea cu un interval, caz in care avem urmatoarele optiuni: sub interval, in interiorul intervalului si deasupra intervalului
 - De regula un atribut numeric este testat de mai multe ori de-a lungul unei cai, de la radacina la frunzele arborelui, fiecare test implicand o alta constanta

Arbori de decizie (6)

- Problema aparte – valorile lipsa
 - Pot fi tratate ca o a treia alternativa (daca se considera o valoare distincta pentru acestea)
 - Pot fi tratate intr-un mod specific – fara a li se asocia o valoare distincta
- Daca atributul testat intr-un nod este nominal -> numarul de copii din nod este egal cu numarul valorilor posibile ale atributului
 - => acelasi atribut nu va fi retestat mai jos in arbore
- Uneori attributele sunt divizate in doua submultimi si exista doua ramuri ale arborelui functie de submultimea caruia ii apartine o anumita valoare
 - Atributul ar mai putea fi testat de-a lungul caii

Arbori de decizie (7)

- Solutie mai sofisticata – impartirea speculativa a instantei in mai multe parti si trimiterea fiecărei parti pe cate o ramura in jos pentru a forma la randul sau subarbori pana la nodurile frunza
 - La divizare se asociaza fiecărei ramuri cate o pondere intre 0 si 1 aleasa proportional cu numarul instantelor care corespund ramurii respective
 - Suma tuturor ponderilor este 1
 - O instanta ponderata poate fi apoi divizata intr-un nod mai jos in arbore
 - Eventual diferitele parti ale instantei pot forma cate un nod frunza si deciziile din aceste noduri frunza trebuie recombinate utilizand ponderile filtrate in frunze

Arbori de decizie (8)

- Pot fi construiti si manual
 - Este necesar un mod corespunzator de vizualizare a datelor a.i. sa se poata lua decizia corecta referitor la cel mai potrivit atribut ce trebuie testat si la cel mai potrivit test

Reguli de clasificare

- O alternativa frecventa la arborii de decizie

IF lacrimare=redusa THEN suportabilitate=nu

IF varsta=tanar AND astigmatism=nu AND lacrimare=normal THEN suportabilitate=usor

IF varsta=pre-presbiotic AND astigmatism=nu AND lacrimare=normal THEN suportabilitate=usor

IF varsta=presbiotic AND diagnostic=miop AND astigmatism=nu THEN suportabilitate=nu

IF diagnostic=hipermetropie AND astigmatism=nu AND lacrimare=normal THEN
suportabilitate=usor

IF diagnostic=miopie AND astigmatism=da AND lacrimare=normal THEN suportabilitate=greu

IF varsta=tanar AND astigmatism=da AND lacrimare=normal THEN suportabilitate=greu

IF varsta=pre-presbiotic AND diagnostic=hipermetropie AND astigmatism=da THEN
suportabilitate=nu

Reguli de clasificare

- o regula de clasificare contine 2 componente:
 - Antecedentul (preconditia) – reprezentat de o serie de teste precum cele din arborii de decizie
 - Cel mai frecvent se prezinta sub forma unei suite de expresii legate prin AND – toate testele date de expresiile elementare trebuie trecute ca sa fie valabila regula
 - Poate avea si o expresie logica mai complexa decat o simpla conjunctie
 - Consecventul (concluzia) – clasa sau clasele care sunt acoperite de regula respectiva
- Cand unei instante i se pot aplica mai multe reguli cu concluzii diferite pot sa apara conflicte

Reguli de clasificare (2)

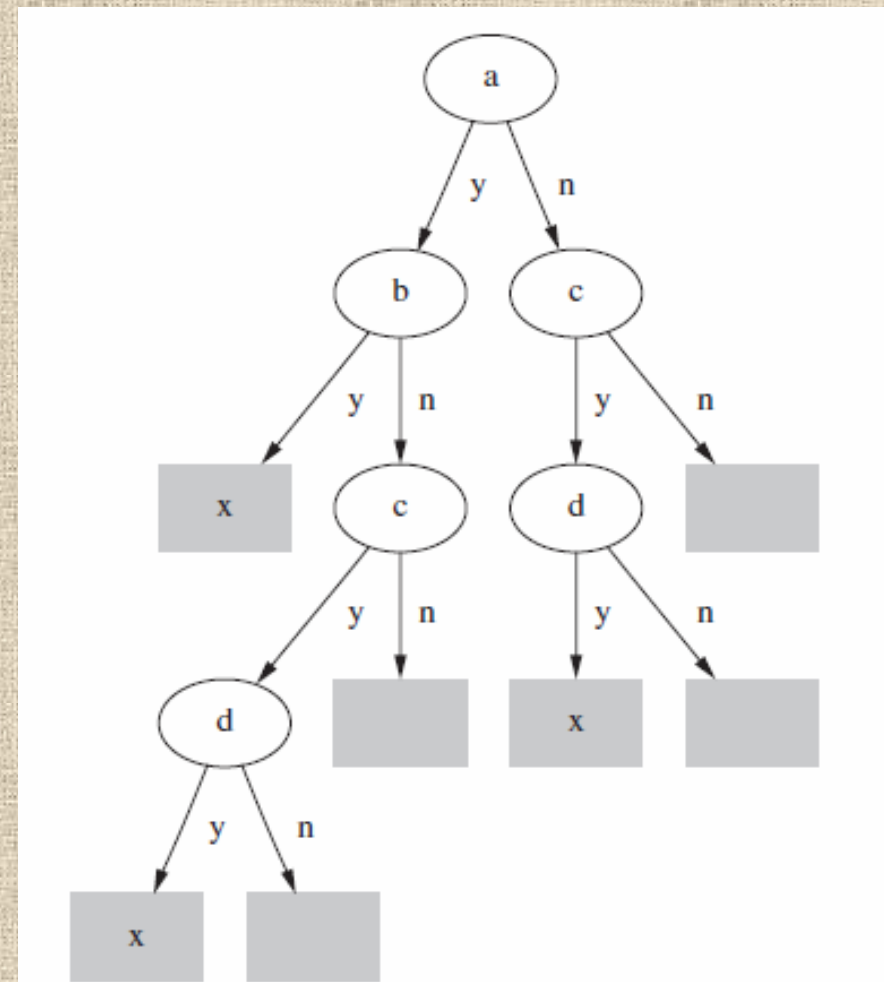
- Este foarte simplu sa se citeasca direct multimea regulilor de clasificare dintr-un arbore de decizie
 - Pentru fiecare frunza este generata cate o regula in care:
 - Antecedentul include conditiile pentru fiecare nod aflat pe calea de la radacina pana la frunza
 - Consecventul este clasa asociata nodului radacina
 - Rezulta reguli clare, indiferent de ordinea in care sunt executate
 - Uneori regulile sunt mai complexe decat este necesar -> pot fi taiate pentru a inlatura testele redundante

Reguli de clasificare vs arbori de decizie

- Arborii de decizie nu pot exprima cu usurinta disjunctiile implicate de diferitele reguli dintr-o multime -> uneori transformarea unei multimi de reguli in arbore de decizie este dificila
- Exemplu:
 - if (a and b) or (c and d) then x
 - if a and b then x
 - if c and d then x

Reguli de clasificare vs arbori de decizie

- Se rupe simetria si se alege un singur test in radacina
- Ex: daca se alege a ca atribut de test atunci regula a doua trebuie repetata de doua ori in arbore
 - => problema subarborilor replicati



Reguli de clasificare vs arbori de decizie

- Fiecare regula poate fi gandita ca un nucleu de cunostinta
 - Pot fi adaugate reguli unui set deja existent, fara a produce vreun neajuns
 - Intr-un arbore de decizie adaugarea unui subarbore ar putea necesita reconstructia intregului arbore
 - Uneori aceasta independenta este iluzorie deoarece ignora modul in care se executa intregul set de reguli
- Daca se presupune ca ordinea interpretarii este aleatoare nu este clar ce trebuie facut cand reguli diferite conduc la concluzii diferite pentru aceeasi instanta
 - Aceasta situatie nu poate sa apara pentru regulile citite direct dintr-un arbore de decizie deoarece redundanta inclusa in structura regulilor previne orice ambiguitate de interpretare

Reguli de clasificare - probleme

- Daca un set de reguli ofera clasificari multiple pentru un anumit exemplu:
 - fie nu se ajunge la nici o concluzie
 - fie se poate evalua cat de des se activeaza fiecare regula in multimea datelor de antrenare si se poate considera cea mai populara dintre reguli
 - Se pot obtine rezultate radical diferite
- Se poate intalni o instanta ce nu poate fi clasificata de nici o regula
 - Fie clasificarea unui astfel de exemplu esueaza
 - Se alege cea mai frecvent intalnita clasa si se atribuie implicit exemplului
 - Se pot obtine rezultate diferite de realitate

Reguli de clasificare - caz particular

- Atunci cand regulile conduc la o clasa booleana (da/nu) si cand se exprima numai regulile care conduc la un anumit rezultat (exemplu “da”)
 - In mod natural se face ipoteza ca daca o instanta nu e in clasa “da” ea trebuie sa fie in cealalta clasa (ipoteza de lume inchisa)
 - In acest caz regulile nu sunt conflictuale si nu dau nastere la interpretari ambigui
- Un astfel de set de reguli poate fi scris ca o expresie logica numita forma normal disjunctiva (disjunctie de conditii conjunctive)
- Este un caz special in care fiecare regula opereaza ca o piesa noua si independenta de informatie care intareste disjunctia
 - Se aplica numai rezultatelor booleene si solicita ipoteza de lume inchisa

Reguli de asociere

- Exprima diferite regularitati continute in setul de date
- Diferă de regulile de clasificare prin aceea ca pot prezice orice atribut sau combinatie de atribute (si nu doar o clasa)
- Nu trebuie obligatoriu folosite ca un set de reguli, ca in cazul regulilor de clasificare, ci pot fi considerate individual
- dintr-un volum relativ redus de date se pot genera foarte multe reguli de asociere diferite => interesul va fi indreptat catre cele care se aplica unui numar suficient de mare de instante si care au o acuratete corespunzatoare relativ la instantele asupra carora se aplica
- Se caracterizeaza prin doua marimi:
 - Suportul (acoperirea) unei reguli de asociere – este numarul instantelor pentru care aceasta regula este corecta
 - De regula se exprima ca procent al acestor instante din totalul instantelor din setul de date
 - Confidenta (acuratetea) numarul instantelor pentru care regula este corecta, exprimata ca proportie din instantele asupra carora se aplica
- **Exemplu:**

If temperatura=scazuta then umiditate=normal

Suportul regulii – numarul de zile in care atat temperatura este scazuta cat si umiditatea normala

Confidenta – proportia zilelor cu temperaturi scazute in care umiditatea este normala
- De obicei se specifica valori minime pentru suport si confidenta si se cauta acele reguli pentru care cele doua marimi au valori superioare celor minime stabilite

Reguli de asociere (2)

- Regulile de asociere care au consecvent multiplu trebuie interpretate cu grija

***IF vant=F AND desfasurare joc=nu THEN
aspect=insorit AND umiditate=mare***

- Regula de mai sus NU este o exprimare scurta a setului de reguli

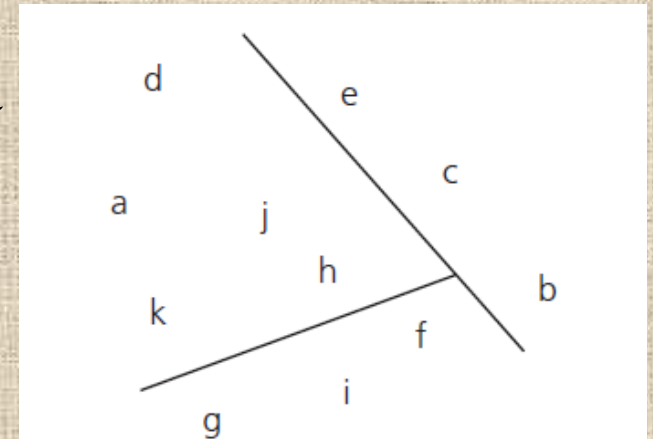
***IF vant=F AND desfasurare joc=nu THEN
aspect=insorit***

***IF vant=F AND desfasurare joc=nu THEN
umiditate=mare***

- Se poate verifica faptul ca acestea din urma satisfac conditiile de suport si confidenta, dar au o alta semnificatie decat regula initiala

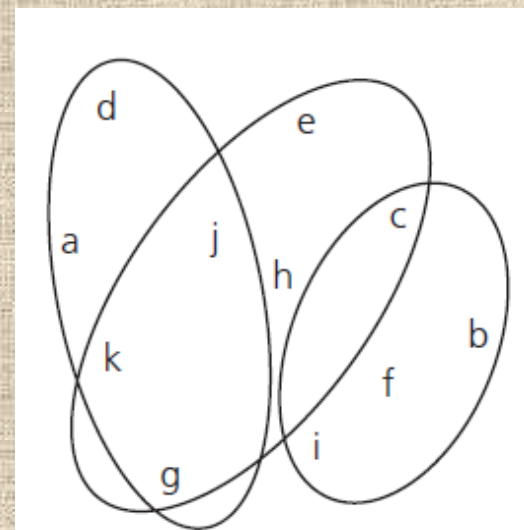
Clusterare

- iesirile iau forma unor diagrame care indica modul in care sunt grupate instantele in cluster
- se pot utiliza mai multe tipuri de algoritmi
- Cel mai simplu caz implica alocarea unui numar de cluster fiecarei instante
 - ar putea fi descris prin plasarea instantelor intr-un spatiu bidimensional si partitionarea acestuia astfel incat sa arate fiecare cluster



Clusterare (2)

- Exista algoritmi care permit unei instante sa apartina mai multor cluster
 - diagrama Venn – plaseaza instantele in doua dimensiuni si se deseneaza subseturi suprapuse reprezentand fiecare cate un cluster
- Asocierea instantelor cu clusterurile se poate face cu o anumita probabilitate



	1	2	3
a	0.4	0.1	0.5
b	0.1	0.8	0.1
c	0.3	0.3	0.4
d	0.1	0.1	0.8
e	0.4	0.2	0.4
f	0.1	0.4	0.5
g	0.7	0.2	0.1
h	0.5	0.4	0.1

Clusterare (3)

- Se poate produce o structura ierarhica de clusterare astfel incat la nivelul de top spatiul instantelor se imparte intr-un anumit numar de clusterare (in general un numar mic) si acestea se divid, la randul lor, in propriile subclusterare, s.a.m.d.
 - Se obtine o *dendograma* (similara unei diagrame arborescente)
- Clustering-ul este adesea umat de o faza de constructie a unui arbore de decizie sau a unui set de reguli cu scopul de a aloca fiecare instanta cluster-ului caruia ii apartine
 - => de cele mai multe ori clusteringul este un pas in procesul de descriere structurala

