

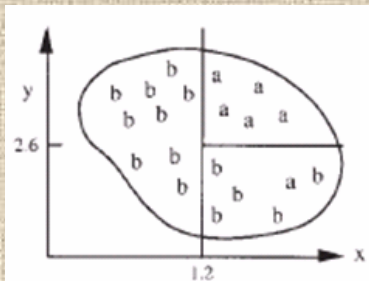
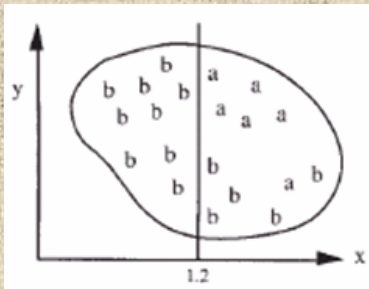
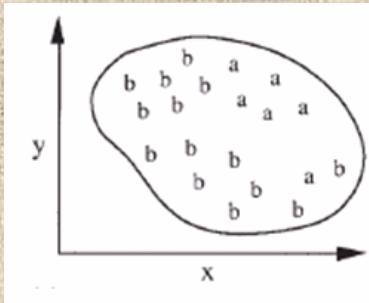
TEHNICI DE EXPLORARE A DATELOR -DATA MINING -

CURS 7

Constructia regulilor

- Alternativa la constructia arborilor de decizie:
 - Se considera fiecare clasa in parte si se cauta o modalitate de a acoperi toate instantele din aceasta clasa si in acelasi timp, de a exclude instantele care nu fac parte din clasa respectiva
 - Aceasta abordare – acoperire (covering)- in fiecare etapa se identifica o regula care “acopera” anumite instante
 - => un set de reguli si nu un arbore de decizie

- Metoda poate fi usor vizualizata intr-un spatiu bidimensional



- Se cauta pentru inceput o regula care sa acopere valorile *a*

If $x > 1,2$ then class=a

- Cu toate acestea regula acopere aproape le fel de multe cazuri *b* ca si cazurile *a*
- => trebuie adugat un nou test care sa divizeze in continuare, intr-un mod potrivit spatiul

If $x > 1,2$ and $y > 2,6$ then class=a

- Se obtine o regula care acopera **aproape** toate valorile *a*
- Daca este necesar ca aceste valori sa fie acoperite **in totalitate**, atunci setul trebuie completat cu inca o regula de genul:

If $x > 1,4$ and $y < 2,6$ then class=a

- Printr-o procedura similara se pot gasi reguli care sa acopere *b*-urile

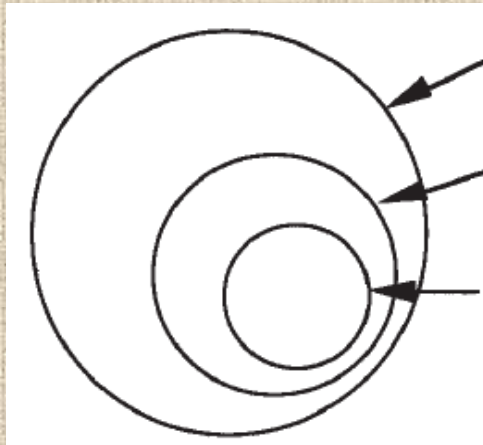
If $x \leq 1,2$ then class=b

If $x > 1,2$ and $y \leq 2,6$ then class=b

Algoritm de acoperire

- Opereaza prin adaugarea de teste unei reguli aflata in constructie, cu intentia de a construi o regula care sa ofere o acuratete maxima
 - diferă de algoritmul divide si cucerește care adaugă teste unui arbore
 - Ambii algoritmi implică găsirea atributului după care se face divizarea
 - Criteriile de găsire a atributului cel mai potrivit sunt diferite:
 - Algoritmul divide si cucerește (ID3) alege un atribut care maximizează câștigul de informație
 - Algoritmul “acoperire” alege o pereche atribut-valoare care maximizeaza probabilitatea clasificarii dorite`
- EXEMPLUL DE LA PAG 107-108

Exemplu



Spatiul exemplelor

Exemple acoperite de o regula in constructie

Exemple acoperite de o regula dupa adaugarea unui nou test

- **Notatii folosite:**
 - C clasele considerate
 - T multimea instantelor de antrenare
 - R regula construita
 - A multimea atributelor
 - v valoarea considerata pentru atributul de test
 - p multimea cazurilor pozitive acoperite de regula
 - t multimea totala a cazurilor pozitive din setul de instante analizat

Pseudocodul algoritmului de acoperire (metoda PRISM)

- Pentru fiecare clasa C
 - Se initializeaza T la multimea instantelor
 - Cat timp T contine instante in clasa C
 - Se creaza o regula R cu antecedent gol (nedeterminat) care prezice clasa C
 - Pana cand R este perfecta (sau nu mai sunt attribute ce se pot folosi) executa
 - Pentru fiecare atribut A nefolosit in R, si fiecare valoare v,
 - » Se adauga conditia A=v in antecedentul regulii R
 - » Se selecteaza A si v a.i. sa maximizeze acuratetea p/t (se alege conditia cu cel mai mare p)
 - » Se adauga A=v regulii R
 - Se elimina instantele acoperite de R din T

Exemplu

Varsta	Prescriptie de ochelari	Astigmatism	Rata de lacrimare	Recomandare lentile de contact
tanar	miopie	nu	reduasa	nu
tanar	miopie	nu	normala	soft
tanar	miopie	da	reduasa	nu
tanar	miopie	da	normala	hard
tanar	hipermetropie	nu	reduasa	nu
tanar	hipermetropie	nu	normala	soft
tanar	hipermetropie	da	reduasa	nu
tanar	hipermetropie	da	normala	hard
pre-presbiopic	miopie	nu	reduasa	nu
pre-presbiopic	miopie	nu	normala	soft
pre-presbiopic	miopie	da	reduasa	nu
pre-presbiopic	miopie	da	normala	hard
pre-presbiopic	hipermetropie	nu	reduasa	nu
pre-presbiopic	hipermetropie	nu	normala	soft
pre-presbiopic	hipermetropie	da	reduasa	nu
pre-presbiopic	hipermetropie	da	normala	nu
presbiopic	miopie	nu	reduasa	nu
presbiopic	miopie	nu	normala	nu
presbiopic	miopie	da	reduasa	nu
presbiopic	miopie	da	normala	hard
presbiopic	hipermetropie	nu	reduasa	nu
presbiopic	hipermetropie	nu	normala	soft
presbiopic	hipermetropie	da	reduasa	nu
presbiopic	hipermetropie	da	normala	nu
presbiopic	hipermetropie	da	normala	nu

If ? then recomandare=hard

- Pentru termenul necunoscut “?” pot exista urmatoarele noua variante:

varsta = tanar 2/8

varsta = pre-presbiopic 1/8

varsta = presbiopic 1/8

prescriptie = miopie 3/12

prescriptie = hipermetropie 1/12

astigmatism = nu 0/12

astigmatism = da 4/12

rata de lacrimare = reduasa 0/12

rata de lacrimare = normala 4/12

Se alege conditia cu rata cea mai mare de cazuri corecte

If astigmatism = da then recomandare=hard

Regula anterioara nu este exacta deoarece acopera numai 4 cazuri corecte si 12 cazuri incorecte

- Se cauta o alta regula de forma

If astigmatism = da and ? then recomandare=hard

Varsta	Prescriptie de ochelari	Astigmatism	Rata de lacrimare	Recomandare lentile de contact
tanar	miopie	da	reduasa	nu
tanar	miopie	da	normala	hard
tanar	hipermetropie	da	reduasa	nu
tanar	hipermetropie	da	normala	hard
pre-presbiopic	miopie	da	reduasa	nu
pre-presbiopic	miopie	da	normala	hard
pre-presbiopic	hipermetropie	da	reduasa	nu
pre-presbiopic	hipermetropie	da	normala	nu
presbiopic	miopie	da	reduasa	nu
presbiopic	miopie	da	normala	hard
presbiopic	hipermetropie	da	reduasa	nu
presbiopic	hipermetropie	da	normala	nu
presbiopic	hipermetropie	da	normala	nu

Pentru termenul necunoscut “?” pot exista urmatoarele sapte variante:

- varsta = tanar 2/4
- varsta = pre-presbyopic 1/4
- varsta = presbyopic 1/4
- prescriptia = miopie 3/6
- prescription = hipermetropie 1/6
- rata de lacrimare = reduasa 0/6
- rata de lacrimare = normala 4/6

Se alege conditia cu rata cea mai mare de cazuri corecte

If astigmatism = da and rata de lacrimare = normala then recomandare = hard

??? Regula este suficient de exacta???

- Se inceara completarea regulii astfel

If astigmatism = da and rata de lacrimare = normala and ? then recomandare=hard

Varsta	Prescriptie de ochelari	Astigmatism	Rata de lacrimare	Recomandare lentile de contact
tanar	miopie	da	normala	hard
tanar	hipermetropie	da	normala	hard
pre-presbiopic	miopie	da	normala	hard
pre-presbiopic	hipermetropie	da	normala	nu
presbiopic	miopie	da	normala	hard
presbiopic	hipermetropie	da	normala	nu

Pentru termenul necunoscut “?” pot exista acum urmatoarele variante:

- varsta = tanar 2/2
- varsta = pre-presbyopic 1/2
- varsta = presbyopic 1/2
- prescriptia = miopie 3/3
- prescription = hipermetropie 1/3

Ar trebui ales acum dintre prima sau cea de-a patra conditie (evaluate ca si raport fiecare conduc la valoarea 1, dar au diferite acoperiri)

If astigmatism = da and rata de lacrimare = normala and prescriptia = miopie then recomandare = hard

regula acopera trei din cele patru cazuri “pozitive”

Cazul 2

- Urmand acelasi rationament se poate considera ca prima conditie de test
If Varsta=tanar then recomandare=hard
- Care va fi regula finala???

Varsta	Prescriptie de ochelari	Astigmatism	Rata de lacrimare	Recomandare lentile de contact
tanar	miopie	nu	redusa	nu
tanar	miopie	nu	normala	soft
tanar	miopie	da	redusa	nu
tanar	miopie	da	normala	hard
tanar	hipermetropie	nu	redusa	nu
tanar	hipermetropie	nu	normala	soft
tanar	hipermetropie	da	redusa	nu
tanar	hipermetropie	da	normala	hard

Pentru termenul necunoscut “?” pot exista urmatoarele sapte variante:

- prescriptia = miopie 1/3
- prescription = hipermetropie 1/3
- astigmatism=da 2/2
- astigmatism=nu 0/4
- rata de lacrimare = redusa 0/4
- rata de lacrimare = normala 2/2

- Se alege

If Varsta=tanar and astigmatism=da then recomandare=hard

Varsta	Prescriptie de ochelari	Astigmatism	Rata de lacrimare	Recomandare lentile de contact
tanar	miopie	nu	redusa	nu
tanar	miopie	nu	normala	soft
tanar	hipermetropie	nu	redusa	nu
tanar	hipermetropie	nu	normala	soft

- se observa ca in setul de exemple ramas nu mai exista cazuri “corecte”
- Este posibila completarea regulii cu cealalta conditie care acopera cazurile din etapa anterioara

If Varsta=tanar and astigmatism=da and rata de lacrimare = normala then recomandare=hard

Implementarea constructiei regulilor de clasificare

- In realitate, problema generarii de reguli este aceea ca tind sa se suprapuna datelor de antrenare si nu pot fi corect generalizate pe seturi de test independente (mai precis pe date care contin zgomote)
 - => este necesar sa existe modalitati de a masura valoarea reala a regulilor individuale
 - abordarea standard de a stabili valoarea unei reguli este aceea de a se evalua ratele de eroare pe un set independent de instante extrase din setul de antrenare

Criteriul de selectie a testului

- In exemplul anterior s-a folosit testul care maximizeaza raportul p/t unde:
 - p = numarul instantelor pozitive
 - t = numarul total de instante pe care le acopera noua regula
 - => intentia de a maximiza corectitudinea regulii pe baza faptului ca, o regula este cu atat mai corecta cu cat acopera o proportie mai ridicata de exemple pozitive
 - Alternativa
 - Calculul unui castig de informatie dupa expresia:
$$P[\log p/t - \log P/T]$$
- Unde:
- P si T sunt numarul de cazuri pozitive si numarul de instante care satisfac regula inainte de adaugarea noului test
- Justificare – expresia data reprezinta informatia totala castigata relativ la exemplele pozitive curente, care este data de numarul acestor exemple pozitive care satisfac noul test

Criteriul de selectie a testului (II)

- Criteriul de baza in alegerea unui test de adaugat unei reguli este legat de gasirea aceuia care acopera cat mai multe exemple pozitive si cat mai putine exemple negative
- 1. euristica bazata pe corectitudine – considera procentul de exemple pozitive din toate exemplele acoperite de regula atinge maximul cand nu exista exemple negative acoperite de regula
 - => un test care face regula “exacta” va fi preferat unuia care o face “inexacta” fara sa tina cont de cate exemple pozitive acopera prima dintre reguli si cate acopera cea de-a doua
 - Exemplu: daca trebuie sa alegem intre un test care copera un exemplu care este pozitiv si unul care acopera 1000 de exemple pozitive si trei negative, se va prefera primul test
- 2 euristica bazata pe informatie se focalizeaza pe acoperirea unui numar mare de exemple pozitive fara a lua in considerare daca regula creata este exacta sau nu

Criteriul de selectie a testului (III)

- Ambii algoritmi continua adaugarea de teste pana se produce o regula finala exata, ceea ce presupune ca regula va fi finalizata mai rapid prin utilizarea masurii corectitudinii, sau vor fi adaugati mai multi termeni daca se utilizeaza masura bazata pe informatie
 - => corectitudinea ar putea gasi cazurile speciale(pe care sa le elimine complet) lasand pentru un moment ulterior gasirea regulilor mai generale
 - Ac lucru devine mai simplu deoarece cazurile speciale au fost deja considerate
 - => masura bazata pe informatie incearca generarea regulilor cu un grad mare de acoperire si lasa pentru un moment ulterior tratarea cazurilor speciale
- Nu exista retete pentru alegerea celei mai bune variante

Valori lipsa si attribute numerice

- In algoritmul de acoperire, **valorile lipsa** pot fi tratate ca si cum nu ar corespunde niciunui test
 - Potrivit cand se priduice o lista de decizii deoarece incurajeaza algoritmul de invatare sa separe sistantele pozitive utilizand teste cunoscute ca fiind reusite
 - Efectul:
 - Fie instantele cu valori lipsa sunt considerate de reguli care implica alte attribute ce nu au valori lipsa
 - Fie orice decizie referitoare la acestea va fi amanata pana cele mai multe din celelalte instante au fost luate in considerare, moment in care testele vor putea implica alte attribute
- => algoritmi de acoperire pentru liste de decizie au avantaje asupra algoritmilor de inductie a arborilor de decizie deoarece exemplele delicate pot fi tratate mai tarziu in proces, cand ele vor putea fi tratate ca fiind mai putin complicate deoarece celelalte exemple au fost in mare parte clasificate si eliminate din multimea instantelor
- **Atributele numerice** sunt tratate ca si in cazul inductiei arborilor de decizie