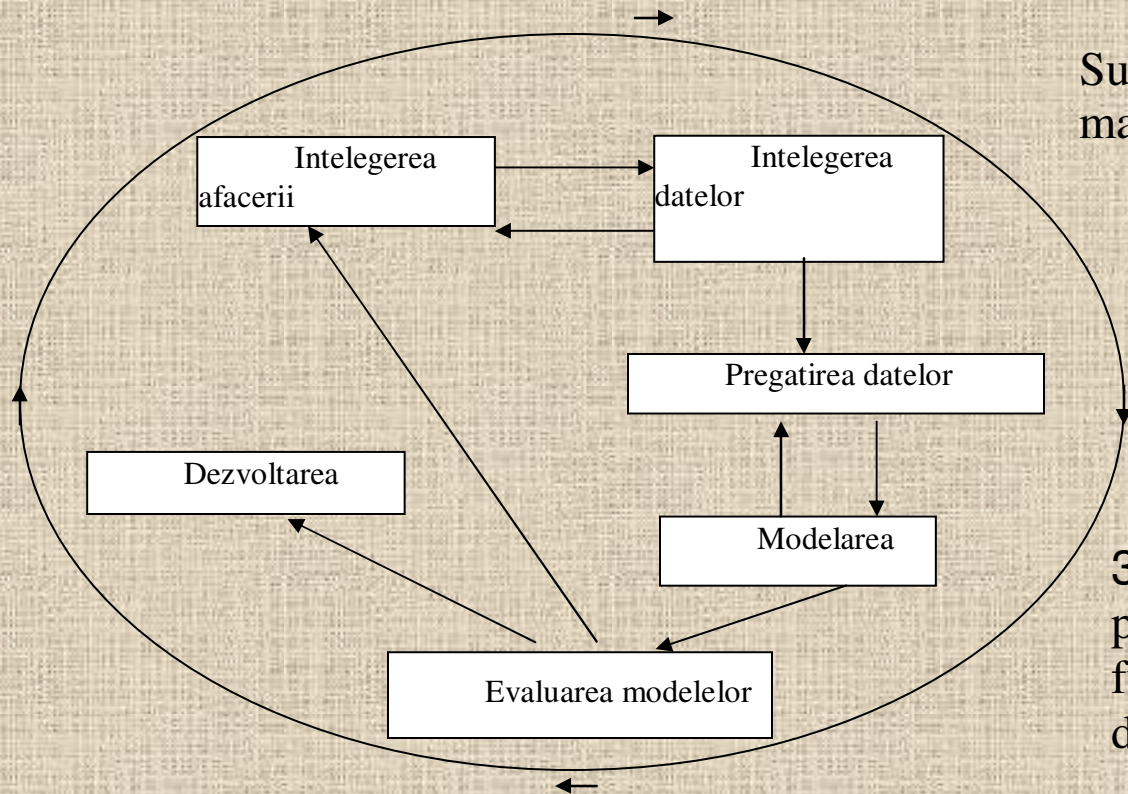


TEHNICI DE EXPLORARE A DATELOR -DATA MINING -

CURS 2

Modelul industrial (CRISP-DM)



Succesiune de pași, parcurși într-o manieră interactivă și iterativă

1. analiza scopurilor declarate de utilizatorul final și primirea tuturor cunoștințelor anterioare necesare.
2. datele țintă sunt pregătite și curățate de tot ceea ce înseamnă zgomote sau valori izolate.

3. se găsește caracteristica utilă pentru reprezentarea datelor, funcție de obiectivul sarcinii de descoperire.

4. se alege și se aplică un anumit algoritm de explorare a datelor în scopul de a prezice valorile viitoare ale variabilelor de interes sau de a găsi tiparele din date, interpretabile de factorul uman..

5. tiparele sunt interpretate și evaluate cu ajutorul unor instrumente specializate, cum ar fi cele de vizualizare.

Atribute, multimii de date si modalități de stocare

- La nivel elementar o singura unitate de informație este o valoare a unei caracteristici (atribut)
 - Fiecare caracteristică poate lua un număr de valori distincte
- Obiectele – descrise prin seturi de caracteristici (*features*) sunt reprezentate sub forma unor mulțimi de date care sunt memorate în fișiere flat, baze de date sau depozite de date (data warehouse)

Valori, caracteristici și obiecte

1. Valori

- In data mining se lucrează cu două tipuri de valori
 - numerice (numere reale, întregi, numere prime, etc...)
 - descriptive (simbolice, categoriale) – se refera la concepte calitative (culori, dimensiuni...)
 - împart instanțele in diferite categorii
 - bărbat, femeie
 - fumător, nefumător
 - căsătorit/necăsătorit/văduv/divorțat.....
 - hipertensiv/hipotensiv
- Uneori datele numerice pot fi tratate ca date categoriale
 - Ex. numărul de mașini deținute împart persoanele respective în categoriile corespunzătoare

Valori, caracteristici si obiecte

2. Caracteristici (attribute)

- Sunt descrise de un set de valori corespunzatoare
 - Funcție de tipul acestor valori pot fi
 - Discrete
 - Continui
- NOTA: cunoașterea tipului valorilor are importanță practică
 - Sunt metode de preprocesare a datelor sau de data mining aplicabile numai datelor descrise prin valori discrete.
 - => este necesar sa se realizeze înaintea acestor etape un proces de discretizare pentru a transforma valorile continui in valori discrete

Valori, caracteristici și obiecte

3. Obiecte

- Entități descrise prin una sau mai multe caracteristici

- înregistrari, exemple, cazuri...

- Dacă un obiect este descris prin:

- mai multe caracteristici – date multivariate

- o singură caracteristică – date univariate

Caracteristica:	
Nume: Daniela Iliescu	simbolica nominala
Sex: femeie	simbolica binara (multimea valorilor: femeie/barbat)
Varsta: 31	numerica, discreta ordinala
TA: 120.0	numerica continua
Colesterol in mg/dl: 320	numerica continua
Numar copii: 3	numerica discreta nominala

- Manipularea diferitelor tipuri de caracteristici și valori – problema importanta in data mining
 - => orice operație asupra mai multor obiecte (compararea caracteristicilor, calculul distanțelor etc.) trebuie atent analizată și proiectată

Mulțimi de date

- grupează obiectele descrise prin aceleași caracteristici
- multe instrumente de data mining consideră mulțimile de date organizate ca fișiere flat în format tabelar cu linii și coloane
 - Liniile -> obiecte
 - Coloanele -> caracteristici
 - => un fișier (text) ce conține un masiv de date bidimensional
 - Este generat din date stocate în alte formate mai complexe: baze de date sau foi de calcul tabelar

Nume	Varsta	Sex	TA	Data masurarii TA	colesterol	Data masurarii
Ion Popescu	67	barbat	120	05/05/2005	null	null
Irina Filip	34	fermeie	130	05/05/2005	332	05/21/2005
Magda Ion	43	Femeie	115	01/03/2007	null	Null
Dan Dumitriu	55	barbat	120	06/02/2006	405	09/09/2007
...

Modalități de stocare a datelor

- Metodele de data mining pot fi aplicate unui număr mare de formate de date
 - baze de date
 - depozite de date
 - www
 - sisteme avansate de baze de date:
 - Baze de date orientate obiect sau baze de date obiect-relationale
 - Baze de date tranzacționale, spațiale, temporale text sau multimedia
- Motive pentru folosirea sistemelor specializate:
 1. Mulțimea de date nu poate fi încărcată odată în întregime în memoria sistemului utilizat pentru data mining => un sistem de gestiunea datelor poate fi utilizat pentru extragerea datelor
 2. Metodele de data mining pot necesita numai anumite subseturi de date => un SGBD poate accesa eficient datele necesare
 3. Datele pot fi adăugate sau actualizate dinamic, de cele mai multe ori de persoane diferite din locatii diferite => un SGBD poate gestiona actualizările concurente si poate oferi facilități de recuperare in caz de defect
 4. Fișierele flat pot conține părți importante de informație redundantă, care pot fi evitate prin stocarea datelor in baze de date

Modalitati de stocare a datelor: baze de date

- SGBD-urile oferă numeroase servicii utile in data mining
 - abilitatea de a:
 - defini structura datelor,
 - de a stoca datele si de a le accesa concurent,
 - de a asigura securitatea si consistenta datelor
- Cel mai comun tip de SGBD-uri întâlnite – SGBD relaționale
 - Caracteristica importanta – existenta SQL care poate furniza:
 - acces eficient la porțiuni ale bazei de date
 - Ex.: se pot extrage informații referitoare la analizele efectuate intr-un interval de timp => lista tuturor pacienților testați – mult mai simplu decât in cazul unui fișier flat
 - posibilitatea de agregare a datelor
 - Ex.: cate analize de un anumit tip au fost efectuate intr-un anumit interval

Ex. baza de date relationala continand aceleasi date ca fisierul flat

Tabelul Pacienti

IDPac	Nume	DDN	sex
1	Ion Popescu	ddn1	barbat
2	Irina Filip	ddn2	femeie
3	Magda Ion	ddn3	femeie
4	Dan Dumitriu	ddn4	barbat

Tabelul Analize

IDAn	Denumire
1	Tensiune arteriala
2	Colesterol

Tabelul Investigatii

IDPac	IDAn	Data	Valoare
1	1	05/05/2005	120
2	1	05/05/2005	130
2	2	05/05/2005	332
3	1	01/03/2007	115
4	1	06/02/2006	120
4	2	09/09/2007	405

Modalitati de stocare a datelor: depozite de date

- Scenariu: exista mai multe clinici, in locatii diferite aparținand aceleiași companii
 - baza de date – asigura posibilitatea analizei datelor dintr-o anumita clinica
 - analiza datelor din toate clinicile – complicată in SGBD
- => mai potrivita utilizarea unui depozit de date (Data Warehouse)
 - Principal scop – furnizarea datelor pentru analize:
 - Datele sunt organizate pe subiecte de interes pentru useri (pacienti, tipuri de teste, diagnostic...)
 - Permite analize din perspectiva istorica
 - Utilizează o structura multidimensionala sub forma unui cub de date
 - OLAP – apelează la cunoștințe fundamentale din domeniu pentru a prezenta datele la diferite niveluri de abstractizare
 - Principalele operatii: roll-up si drill- down

Modalitati avansate de stocare a datelor

- Au aparut noi tipuri de date:
 - date tranzacționale
 - date spatiale
 - hypertext (HTML, XML)
 - date multimedia (combinații de text, imagini, înregistrări video si audio)
 - date temporale
 - => sunt necesare baze de date specializate care utilizează structuri de date si metode eficiente de manipulare a acestor structuri complexe de date
 - Provocarea – lucrul cu:
 - obiecte de dimensiuni variabile,
 - cu date structurate si semistructurate,
 - text nestructurat de dimensiuni variabile,
 - formate de date multimedia
 - **volume foarte mari de date**

Baze de date orientate obiect

- Se bazează pe paradigma programarii orientate obiect
 - Tratează fiecare entitate stocată ca pe un obiect care încapsulează:
 - Un set de variabile prin care este descris
 - Un set de mesaje utilizate pentru a comunica cu alte obiecte
 - Un set de metode care conțin cod ce implementează mesajele
 - Exemplu:
 - obiectul – pacientul
 - Variabile: nume, adresa, sex...
 - Metode: cod ce poate implementa mesaje precum” Sa se găsească valorile anumitor analize la anumite momente de timp”.
- Elementul cheie – abilitatea de a grupa obiectele similare sau identice în clase, care, la rândul lor, pot fi organizate în ierarhii

Baze de date obiect-relationale

- Sunt construite pe modelul de date obiect-relațional
 - O bază de date relațională extinsă prin furnizarea unui set de tipuri complexe de date ce permit manipularea obiectelor complexe
 - => necesitatea utilizării unui limbaj de interogare specializat, capabil să regăsească date complexe din baze de date

Baze de date tranzactionale

- Sunt stocate in fișiere flat si constau in inregistrari care reprezintă tranzacții
 - Structura unei tranzactii include un identificator unic și o mulțime de itemi (articole)
- Ex. – inregistrarea cumpărăturilor dintr-un magazin
- Eventualele informatii aditionale sunt stocate in fisiere separate si pot include: numele clientului, numele casierului, data tranzactiei, magazinul, etc.
- Diferentele dintre b.d. relationale si b.d. tranzactionale:
 - b.d.t. memoreaza o multime de articole si nu o multime de valori ale unor caracteristici interconectate
 - b.d.t. memoreaza informatii relativ la prezenta/absenta unui item in timp ce b.d.r. memoreaza date despre valorile caracteristicilor pe care le posedă un item (exemplu, entitate)
- Ex. O b.d.t. memoreaza numele analizelor efectuate, ca itemi
 - dacă fiecare tranzactie reprezinta o singura vizita la clinica a unui pacient se pot determina ce analize se efectueaza frecvent impreuna
 - pe baza acestor rezultate se poate adapta practica medicala

Baze de date spatiale

- Sunt proiectate sa gestioneze date cu caracter spatial:
 - harti geografice
 - imagini de la sateliti
 - imagini medicale
 - Datele spatiale se pot reprezenta in 2 moduri:
 - în format raster – harti de pixeli n-dimensionale
 - în format vectorial – toate obiectele sunt reprezentate ca obiecte geometrice simple (linii, triunghiuri, poligoane) si se foloseste geometria vectoriala pentru a calcula relatiile dintre obiecte
- Permit userului sa obtina informatii referitoare la datele stocate
 - Ex. daca pacientii care locuiesc intr-o anumita vecinatate sunt mai expusi unor riscuri

Baze de date temporale

- “baze de date serii de timp” – stochează date legate de diferite momente in timp
 - extind bazele de date relaționale pentru a manipula caracteristica timp
 - attributele temporale pot fi definite utilizând mărci de timp pentru diferite noțiuni, precum zile și luni, ore și minute, zile ale săptămânii, etc
 - baza de date păstrează caracteristicile prin stocarea secvențelor valorilor acestora care se modifica in timp
 - Diferă de o baza de date relațională care memorează cele mai recente valori ale unor caracteristici
 - Permit userului să găsească tipare în date, care, de cele mai multe ori, evaluează tendințe de schimbare
 - Ex. Se poate urmări dacă TA are tendințe crescătoare sau descrescătoare odată cu înaintarea în vârstă a unui pacient sau a unui grup cu caracteristici similare (colesterol mare)

Baze de date multimedia

- permit stocarea, regăsirea si manipularea imaginilor si inregistrarilor audio sau video
- datorita dimensiunilor foarte mari ale surselor de date este necesar sa se utilizeze medii de stocare specializate si tehnici de cautare specifice
 - Ex. Gasirea relatiilor intre inregistrarea video cu miscarile inimii si inregistrarea batailor inimii

Relația dintre valori, caracteristici, obiecte multimi de date si modalitati de stocare

