Cap. 14. Data mining

Descoperirea de cunostințe din bazele de date (Knowledge Discovery in Databases – KDD) sau **extragerea de date (Data Mining – DM)** reprezintă efortul de a înțelege, analiza și eventual de a utiliza o cantitate imensă de date disponibile

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Cap. 11. Data mining

Definiţii

Definiţie

Data Mining este procesul descoperirii (semi)automate a informațiilor utile în depozite mari de date (Tan et al).

Definiţie

Data Mining este analiza seturilor de date – deseori de dimensiuni mari – rezultate prin observații pentru a găsi relații noi și pentru sumarizarea datelor în moduri care sunt atât ușor de înțeles cât și utile celui ce deține datele (Hand et al).

Definiţie

Data mining este procesul netrivial de extragere a informației implicite, anterior necunoscute, interesante și potențial utile din date, de regulă sub forma de modele și șabloane de cunoaștere (Schapiro et al).

Data mining

Termeni alternativi:

- mineritul *cunoștințelor* din date
- extragere de cunoștințe (eng: Knowledge Discovery) – sinonim discutabil
- analiza date/șabloane

Ce NU e Data Mining:

- găsirea datelor complete privind o persoană folosind interogare într–o bază de date;
- găsirea paginilor web care conțin anumiți termeni;

Acestea sunt activități de *regăsire a informației*.

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

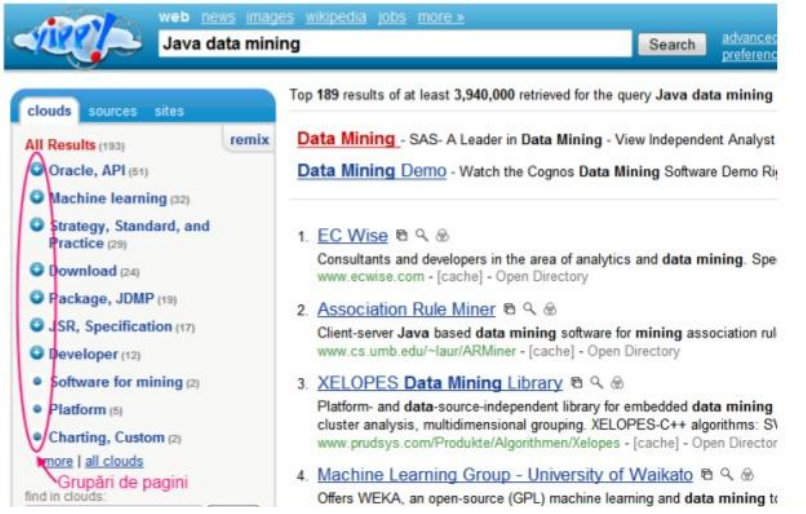
Data mining

Ce poate fi Data Mining:

- să descoperi că anumite nume sunt mai frecvente în unele zone: O'Brien, O'Rurke, O'Reilly în zona Boston;
- gruparea clienților pe baza unui profil de consum comun;
- gruparea paginilor dintr-un motor de căutare pe baza similarităților: motorul search.yippi.com;
- predicția evoluției preturilor la bilete de avion: farecast.com.

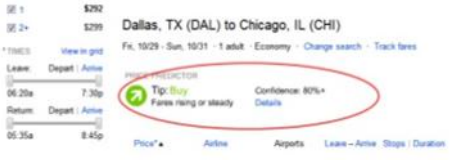
Exemple: Yippi=research-oriented search engine

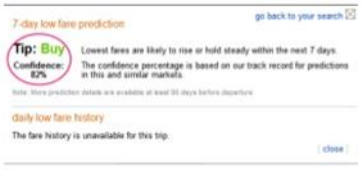
Clustering de pagini web in Yippi



Exemple: Farecast

Farecast: să cumpăr sau nu acum un bilet de avion?



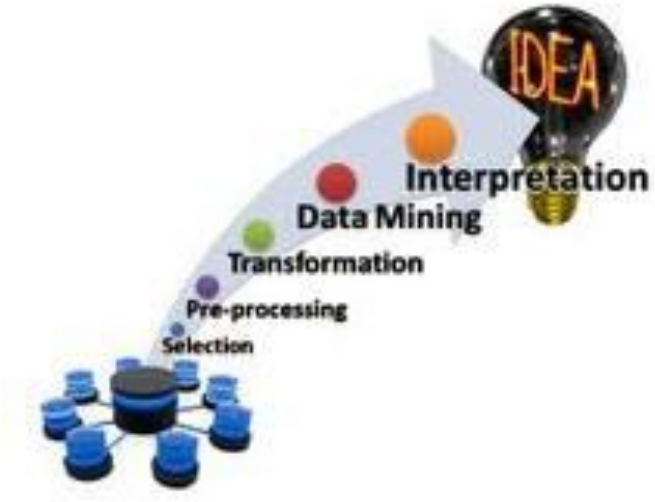


Data mining

De ce Data Mining: din punctul de vedere al afacerilor (1)

- O mulţime de date sunt colectate şi depozitate prin sisteme de data warehouse
  - date din Web, comerţ electronic
  - cumpărături în magazine/lanţuri de desfacere
  - tranzacţii financiare, carduri de debit/credit
- Calculatoarele au devenit tot mai ieftine şi mai puternice; procesarea distribuită este ceva comun.

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ



Data mining

De ce Data Mining: din punctul de vedere al afacerilor (2)

- Presiunea impusă de competiție este motivantă: aducerea unui nou client într–o rețea de telefonie este de până la 4 ori mai scumpă decât păstrarea lui: Customer attrition
- Cerințe specifice mediului de afaceri: customer profiling, targetted marketing, fraud detection
- Probleme stringente: "Care sunt cei mai profitabili clienți?", "Care produse cumpărate atrag achiziția altor produse?", "Care va fi evoluția companiei/pieței pe segmentul ...?", "Care sunt nișele de piață?"

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Data mining

De ce Data Mining: din punct de vedere științific

- În domenii precum medicina, inginerie și știință se acumulează rapid date ce trebuie exploatate pentru a duce la noi descoperiri;
- Exemplu: dezvoltarea de sisteme de sateliți pentru observații climatice;
- Date genetice generate prin "microarrays"; se dorește decodificarea completă a genomului uman, determinarea genelor care cauzează diferite afecțiuni, înțelegerea structurii și funcționalității genelor;
- DM e unealtă de bază pentru bioinformatică = "aplicarea statisticii și a informaticii în domeniul biologiei moleculare".

Knowledge Discovery and Data Mining

Competiții

KDD Cup

- 2015: (Sydney) analysis of social and information networks, mining the web graph, study of cascading behaviors in networks, and the development of algorithmic models of human behavior
- 2012: User Modeling based on Microblog Data and Search Click Data
- 2011: Recomandare de muzică
- 2010: Evaluarea performanțelor studenților
- 2009: Predicția relației cu clienții
- 2008: Cancer de sân
- 2007: Netflix prize
- 2006: embolism pulmonar din date tomografice
- 2005: clasificarea interogărilor de Internet
- 2004: fizica particulelor și biochimie
- 2003: mineritul rețelelor și analiza log-urilor
- competiția merge până în 1997

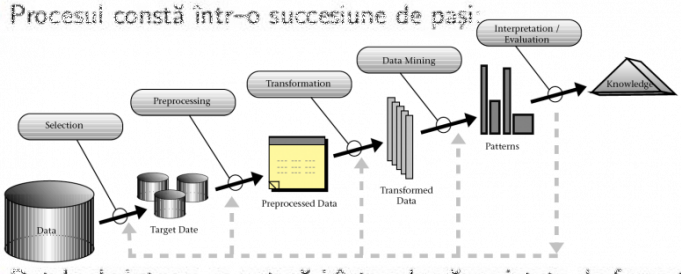
- Alte competiții — www.kdnuggets.com

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Pașii unui proces de extragere de cunoștințe (1)

- Data Mining este parte integrantă a domeniului Knowledge discovery in databases (KDD), care e un întreg proces de conversie a datelor primare în cunoștințe (informație).
- Procesul constă într-o succesiune de pași:



- Datele de intrare se pot găsi într-o largă varietate de formate: fișiere text, baze de date relaționale, date semistructurate (e.g. XML, HTML), imagini, filme etc.

Knowledge Discovery and Data Mining

Pașii unui proces de extragere de cunoștințe (2)

- Datele se selectează din multitudinea de surse;
- Preprocesarea și transformarea pot include: selectarea dimensiunilor, reducerea dimensionalității, tratarea datelor incomplete, normalizarea;
- Preprocesarea și transformarea pot lua chiar și 60% din durata totală a unui proces de extragere a cunoștințelor;
- Partea de Data Mining se face printr–o varietate de tehnici; deseori se testează mai multe metode;
- La final, cunoștințele rezultate sunt post–procesate (*e.g.* se elimină rezultatele invalide sau neinteresante) și trebuie prezentate într–o formă inteligibilă factorilor de decizie (*e.g.* vizualizare sau reguli de forma "if–then"), sau integrate în alte sisteme (*e.g.* sistemele utilizate pentru detectare de fraude);

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Atenție la ce se obține

- Tehnici folosite la preprocesare: testarea ipotezelor prin metode statistice – se elimină rezultatele nerealiste;
- Eliminarea cunoștințelor "neinteresante" — element subiectiv, dependent de cunoștințele anterioare;
- Limitarea complexității modelelor folosite în procesul de DM: "If you torture the data long enough, it will confess" (Ronald Harry Coase, economist);
- Principiul lui Bonferroni: if you look harder than the quantity of data supports, you will find a pattern that "fits".

Knowledge Discovery and Data Mining

Scalabilitatea și dimensiunea datelor

- seturile de date ajung ușor la dimensiuni de giga/tera/peta–bytes;
- France Telecom are o bază de date folosită pentru luarea deciziilor de 30 TB
- Wal-Mart are 20 de milioane de tranzacții pe zi;
- 16 telescoape europene produc 1 Gb pe secundă;
- proiectul genomului uman: 3.4 miliarde de perechi și între 20000 și 25000 gene;
- problemă de descoperire de medicamente: 100000 de atribute; stabilirea reputației URL-urilor: 3231961 de atribute
- Experimentul "Compact Muon Solenoid" la CERN's Large Hadron Collider generează 40 de terabytes de date pe secundă.

 Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Scalabilitatea și dimensiunea datelor (2)

- variante: structuri de date specifice, care să ușureze interogarea datelor
- scalarea pe orizontală sau pe verticală a resurselor hardware;
- scalarea pe verticală: rareori suficientă, datele nu încap în RAM
- scalarea pe orizontală – cazuri remarcabile: Apache Hadoop, Apache Mahout — proiecte open–source.

Knowledge Discovery and Data Mining

Date eterogene și complexe

- atribute eterogene: numerice, categoriale;
- ce faci cu datele lipsă? eliminarea înregistrărilor cu goluri de date nu e întotdeauna o opțiune;
- colecții de documente (*e.g.* pagini Web); date ADN cu structură spațială și secvențială; serii de timp
- tehnicile de DM trebuie să ia în considerare relațiile dintre date (corelație spațială și temporală; conectivitate de grafuri; relație părinte–copil).

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Gestiunea și distribuirea datelor

- datele pot fi prezente în locații multiple, nu doar într–o organizație;
- necesitate: DM distribuit sau suport de tip Data Warehouse
- în caz de distribuire: comunicarea necesară poate să domine timpul de calcul
- în caz de data warehouse: integrarea datelor necesită timp îndelungat
- "data privacy": problemă delicată, diferite aspecte legislative pot interveni

Knowledge Discovery and Data Mining

Analiză nestandard

- Statistica: enuntarea de ipoteze și apoi testarea lor;
- Problemă evidentă: procesul este laborios
- DM are ca scop tocmai determinarea *pe cât posibil* automată a astfel de ipoteze;
- În timp ce statistica este în mare măsură tributară modelelor parametrice, datele reale pot avea cu totul alte distribuții decât cele presupuse;
- Dar statistica oferă unelte utile – de exemplu metode de testare, determinarea intervalelor de confidență, inferența statistică etc.

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Originile DM

- Statistică – eșantionare, estimare, testarea ipotezelor, modele parametrice;
- Inteligență artificială — tehnici de raționament probabilist și management al incertitudinii
- Învățare automată (machine learning) — pornind de la date se creează modele adecvate
- Recunoaștere de șabloane (pattern recognition)
- Sisteme de baze de date – suport pentru stocarea (eventual distribuită a) datelor; probleme pot apărea din cauză că nu toate datele se pot reprezenta ușor sub model relațional;
- Calcul paralel—distribuit — pentru a rezolva problema scalabilității aplicațiilor de DM;

Knowledge Discovery and Data Mining

Sunt două categorii majore de aplicații:

Predicția — scopul e de a prezice valoarea concretă a unui atribut pe baza altor atribute. Atributul ce urmează a fi prezis se numește *variabilă dependentă* sau *țintă*; cele care se folosesc pentru predicție sunt *variabile independente* sau *explicative*;

Descrierea — determinarea de șabloane, *e.g.* corelații, tendințe, grupări, traiectorii, anomalii

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

- Clasificare — predicție
- Grupare (Clustering) — descriere
- Determinarea relațiilor de asociere — descriere
- Descoperirea șabloanelor secvențiale — descriere
- Regresie — predicție
- Detectarea deviațiilor — predicție

Knowledge Discovery and Data Mining

Clasificarea: definiție

- Se pleacă de la o colecție de înregistrări = setul de antrenare
- Fiecare înregistrare e formată din atribute, dintre care unul este "clasa": bun/rau, risc mare/risc moderat/risc mic;
- Scopul este găsirea unui model (a unui mecanism, a unei funcții) care să determine clasa pe baza atributelor;
- Modelul trebuie să facă o clasificare cât mai fidelă pentru înregistrări care nu fac parte din setul de test = date din setul de testare;

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Clasificarea: aplicația 1

Marketing direct:

- scopul: reducerea costurilor de trimitere a reclamelor prin poștă prin alegerea unui set de consumatori pentru care șansele de achiziție a unui produs sunt mari
- modalitate de lucru:
  - se pleacă de la produse similare
  - pentru aceste produse știm dacă au fost sau nu cumpărate de către consumatorii în cauză; asta dă clasa unei înregistrări, ca valoare posibilă din mulțimea {a cumpărat, nu a cumpărat}
  - se colectează date demografice despre clienți, istoricul tranzacțiilor etc.
  - se folosesc aceste date pentru a construi un clasificator.

Knowledge Discovery and Data Mining

Clasificarea: aplicația 2

Prevenirea migrării clientului:

- Scop: să se determine dacă un client al serviciilor oferite este pe cale de a pleca la un competitor
- modalitate de lucru:
  - se folosesc înregistrări detaliate despre tranzacțiile făcute de client (*e.g.* telefonie: apelurile efectuate, rețelele către care s–au efectuat, durata, frecvența);
  - se folosesc date demografice: situația financiară, starea civilă etc.
  - se etichetează clientul ca fiind loial sau nu
  - plecând de la acest set de antrenare se creează un clasificator care să fie utilizat pentru alți clienți

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Clasificarea: aplicația 3

Clasificarea obiectelor cerești

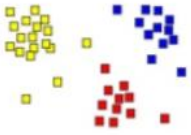
- Scop: să se prezică clasa unor obiecte cerești pe baza imaginilor luate de telescoape
- modalitate de lucru:
  - se pleacă de la o colecție de imagini; caz concret: 3000 imagini cu 23040 x 23040 pixeli pe imagine
  - se segmentează imaginea
  - se măsoară anumite trăsături
  - se construiește un clasificator plecând de la aceste segmente de imagini cu clase atașate - pentru fiecare segment se știe exact ce reprezintă
  - poveste de succes: s–au găsit 16 noi quasari, elemente greu de descoperit și catalogat prin mijloace tradiționale.

    (quasar (quasi-stellar radio source) este un nucleu galactic activ îndepărtat, care emite enorme cantități de energie).

Knowledge Discovery and Data Mining

Clustering: definiție

- Dându–se un set de puncte, fiecare având un set de atribute și o măsură de similaritate, să se găsească grupări (cluster–e) cu proprietatea:
  - punctele care aparțin unui aceluiași cluster sunt similare între ele
  - punctele din clustere separate sunt mai puțin similare
- măsură de similaritate: distanța Euclidiană sau alte măsuri specifice
- deosebire față de clasificare: printre atributele considerate nu există un atribut de clasă



Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Clustering: exemplu

Gruparea automată de documente

- scop: găsirea grupurilor de documente care sunt similare pe baza termenilor pe care îi conțin
- modalitate de lucru
  - se contorizează cuvintele
  - se formează o măsură de similaritate între documente pe baza frecvențelor
  - pe baza similarității se formează grupurile
  - utilitate: pentru un nou document se descoperă rapid care este clusterul căruia îi aparține în mod natural;
- utilitate: detectare de plagiate, căutare de documente similare etc.

Knowledge Discovery and Data Mining

Analiza asocierilor: definiție

- Dându–se un set de colecții de înregistrări, să se producă regulile de dependență care prezic apariția unui item pe baza apariției altor itemi

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Analiza asocierilor: exemple

- găsirea grupurilor de gene care au funcții înrudite
- identificarea paginilor Web dintr–un site care sunt accesate împreună
- Market Basket Analysis: care sunt produsele care se vând bine împreună; în funcție de aceste grupări se poate specula partea de cross-selling (ieftinești un produs dar îl scumpești pe un altul) sau dispunerea pe raft a lor (cele care se vând împreună să fie dispuse apropiat);
- echiparea mașinilor care participă la reparații cu anumite unelte, pentru a reduce numărul de deplasări la client

Knowledge Discovery and Data Mining

Regresie: definiție, exemple

- Prezicerea unui atribut continuu pe baza unor atribute independente;
- Similar cu clasificarea, dar la regresie valorile variabilei dependente sunt numerice
- Intens studiată în statistică și rețele neurale artificiale
- Exemple:
  - prezicerea volumului de vânzări
  - prezicerea vitezei vântului pe baza umidității, presiunii, temperaturii etc.
  - prezicerea consumului de curent într–o anumită perioadă, pe o zonă specificată

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Knowledge Discovery and Data Mining

Detectarea anomaliilor

- detectarea deviațiilor semnificative de la comportamentul normal
- aplicații:
  - detectarea fraudelor cu card bancar
  - detectarea intruziunilor în rețele de calculatoare

Knowledge Discovery and Data Mining

METODE CLASICE DE DATA MINING

- Metode statistice (regresie, Modele lineare generalizate , Arborii de regresie , Analiza variabilității , Modele cu efect mixt , Seriile de timp, etc)
- Vecini (Algoritmul celor mai apropiați k vecini)
- Clustering

TEHNICI DE NOUA GENERAȚIE

- Arbori, rețele neuronale
- Arbori de decizie(Algoritmii CART și CHAID)

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Reprezentarea datelor

Tipuri de date (1)

- Un set de date este o colecție de obiecte-dată (eng: data objects) și de atribute
- Sinonime pentru obiecte-dată: **înregistrare**, **punct**, **vector**, **pattern** (termen ce poate induce confuzie), **eveniment**, **caz**, **eșantion** (termen ce poate induce confuzie), **observație**, **entitate**.
- Obiectele sunt descrise prin **atribute**
- Sinonime pentru atribut: **variabilă**, **caracteristică**, **trăsătură** (feature), **dimensiune** (a nu se confunda cu omonimul din algebră).

Atribute

Date

Id	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No

Reprezentarea datelor

Tipuri de date (2)

Definiție

Atribut: proprietate sau caracteristică a unui obiect ce poate sa varieze fie de la un obiect la altul, fie de la un moment de timp la altul.

- Exemple: culoarea ochilor, temperatura.
- Trebuie făcută diferența între proprietățile atributelor și proprietatile valorilor atributelor:
  - același atribut poate avea valori diferite: înălțimea poate fi măsurată în metri sau picioare
  - diferite atribute pot fi măsurate cu acelasi tip de date, dar proprietățile atributelor pot fi diferite: pentru niște persoane, atributele "inaltime" și "id" sunt reprezentate prin numere întregi; în timp ce are sens sa faci media înălțimilor, nu are nicio noimă media id-urilor; operațiile ce se pot face pentru înaltime (medie, max etc.) nu se aplică și pentru id-uri; id-urile nu au un maxim, în timp ce înălțimea - da.

 Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Reprezentarea datelor

Tipuri de atribute

Există diferite tipuri de atribute:

- Categoriale (calitative)
  - nominale: valori diferite care permit recunoașterea diferențelor; exemple: cod postal, id-uri, culoarea ochilor, genul; operatii permisibile: $=, \neq$;
  - ordinale: valorile permit ordonarea obiectelor; exemple: scara durității mineralelor, grade (militare etc.), numere de imobile; operații permisibile: $=, \neq, <, >$; funcții aplicabile: mediana, percentile etc.
- Numerice (cantitative)
  - interval: se poate face diferența între valori (i.e. există unități de masură asociate); exemple: date calendaristice, temperaturi in grade Celsius sau Fahrenheit; pe lângă operațiile de mai sus admit și adunare, scădere; funcții aplicabile: media, deviatia standard, corelația
  - multiplicabile: permit împărțiri și înmultiri; exemple: temperatura în Kelvin, cantități monetare, număr de elemente, vârsta, greutate; operatii: cele de mai sus și $*$, $/$; funcții aplicabile: media geometrică, variație procentuală.

Reprezentarea datelor

Descrierea atributelor prin numarul de valori (1)

- Atribute discrete:
  - o mulțime cel mult numărabilă de valori;
  - exemple: coduri poștale, cuvinte într-un document
  - se reprezintă cel mai frecvent ca numere naturale
  - caz special — atribute binare: {prezent, absent}
- Atribute continue:
  - valorile sunt exprimate prin numere reale
  - exemple: temperatura, masa
  - dpdv practic reprezentarea se face cu o precizie finită
  - reprezentare actuală: valori în virgulă mobilă

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Reprezentarea datelor

Descrierea atributelor prin numarul de valori (2)

- Valori asimetrice:
  - doar prezența unei trăsături (*i.e.* valoare non–zero) este importantă
  - exemple: vectorul care reprezintă daca niște cuvinte sunt prezente (eventual: de cate ori) într-un document
  - dacă se iau in considerare doi astfel de vectori, contează mai mult cuvintele pe care le au în comun decât cuvintele care lipsesc din ambele documente, simultan

Reprezentarea datelor

Tipuri de *seturi* de date

- Seturi de date de tip: înregistrare, de tip graf și de tip secvență
- Caracteteristici generale:
  - **dimensionalitatea** = numărul de atribute pe care obiectele-dată le au. Un numar de dimensiuni prea mare duce la "*blestemul dimensionalității*"; pentru multe dimensiuni se pot aplica *tehnici de reducere a numarului de dimensiuni*;
  - **caracterul rarefiat al datelor** = procentul de date utile; de exemplu, pentru date asimetrice este numarul de valori nenule. Specularea acestui caracter poate reduce drastic necesarul de memorie sau timpul de calcul;
  - **rezoluția** = scara la care se face raportarea valorilor; e posibil ca scări diferite sa releve (sau sa ascundă) pattern-uri; ex: măsurători meteo raportate pe zile pot arăta iminența unei furtuni, dar la scală de săptămâni așa ceva nu mai e vizibil.

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Reprezentarea datelor

Seturi de date de tip înregistrare

- cel mai des furnizate si frecvent utilizate in aplicatii: multime de obiecte cu un set predefinit de atribute
- nu exista legatura intre inregistrari distincte
- stocare: fisiere text (*e.g.* CSV), Excel, baze de date relationale – views

Atribute

Date

Id	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No

Reprezentarea datelor

Cazuri remarcabile de seturi de date inregistrare (1)

Tranzactii, date specifice cosurilor de cumparaturi:

- exemplu: intr-un magazin, setul de produse cumparate de un client in timpul unei sesiuni de cumparaturi = continutul cosului de cumparaturi
- se analizeaza asocierea intre produsele individuale din tranzactii
- posibilitate de reprezentare: indicator boolean care arata daca un produs anume face sau nu parte dintr-un cos de cumparaturi
- variatie: cate exemplare din produs au fost achizitionate (0, 1, . . .)

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Reprezentarea datelor

Cazuri remarcabile de seturi de date inregistrare (2)

Matrice de date:

- pentru cazul in care datele au acelasi set fix de atribute **numerice**
- fiecare data in parte poate fi considerata un punct in spatiu multidimensional
- fiecare atribut considerat este o dimensiune
- este tipul de date standard pentru analiza statistica
- nota: intre conceptul de dimensiune asa cum e definit in matematica si cel de dimensiune—atribut pot exista diferente

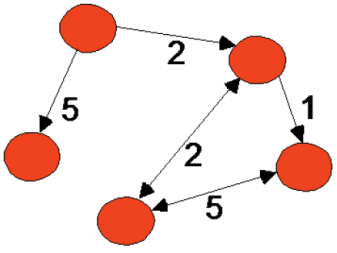
Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

Reprezentarea datelor

Seturi de date de tip graf

- reprezentare convenabilă pentru cazurile:
  - graful reprezintă relații între obiecte
  - înșeși obiectele sunt reprezentate ca graf

exemplu de algoritm ce folosește structura de graf:
algoritmul PageRank




```
<a href="papers/papers.html#bbbb">
Data Mining </a>
<li>
<a href="papers/papers.html#aaaa">
Graph Partitioning </a>
<li>
<a href="papers/papers.html#aaaa">
Parallel Solution of Sparse Linear System of Equations </a>
<li>
<a href="papers/papers.html#ffff">
N-Body Computation and Dense Linear System Solvers
```

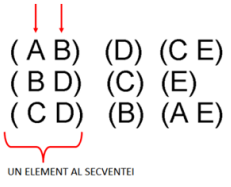
Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Reprezentarea datelor

Seturi de tip secvență

- numite și **date temporale**
- fiecare înregistrare are un atribut suplimentar de timp asociat

- atributele au relații care implică ordonare în timp sau spațiu
- subtipuri: date secvențiale, secvență, serii de timp și date spațiale



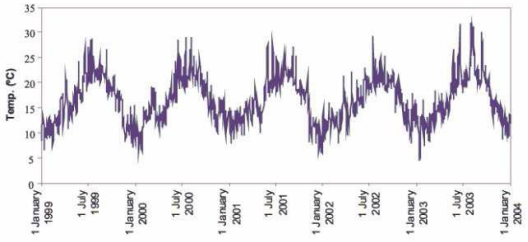
Reprezentarea datelor

Seturi de tip secvență

Ex. informatia genetica

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Ex. date meteo masurate lunar

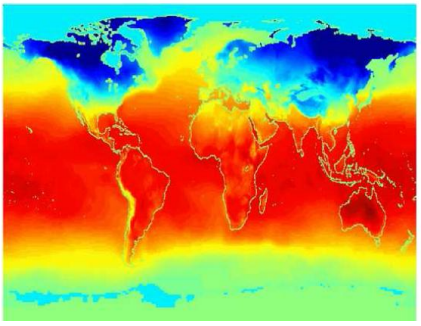


Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Reprezentarea datelor

Seturi de tip secvență: date spațiale

- cazul datelor care au atribute spațiale sau areale
- exemplu: date climatice raportate pe regiuni
- exemplu: date adunate pentru scurgerea unui fluid — poziția diferitelor puncte este inregistrată



Preprocesarea datelor

Scopul pasului de preprocesare

- Strategii si tehnici complexe, ce pot cere până la 60% din timpul total al procesului de extragere de cunoștințe
- Două variante:
  1. selectara obiectelor-dată și a atributelor
  2. crearea/schimbarea de atribute
- Variante de preprocesare:
  - agregare
  - eșantionare
  - reducerea dimensionalității
  - selectarea unui subset de atribute
  - crearea de atribute
  - discretizare și binarizare
  - transformarea variabilelor

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Explorarea datelor

Explorarea datelor

- Explorarea datelor reprezintă investigarea preliminară a datelor, cu scopul de a obține o înțelegere a caracteristicilor lor
- Pasul de explorare poate fi de folos în alegerea pașilor de preprocesare sau analiză
- Se poate folosi abilitatea naturală a oamenilor de a recunoaște pattern-uri
- Domeniul a fost introdus de către statisticianul John Tukey: *Exploratory Data Analysis*, Addison-Wesley
- AED este domeniu opus lui "Confirmatory Data Analysis", care are ca scop testarea ipotezelor statistice, calculul intervalelor de încredere etc.

Vizualizarea datelor

Vizualizare

- Scopul vizualizării: reprezentarea informației într–un mod tabular sau grafic
- Caracteristicile datelor și relațiile dintre elemente pot fi analizate sau raportate
- Calități:
  - oamenii au o abilitate naturală de analiză pentru cantități mari de date prezentate vizual
  - oamenii pot detecta relativ ușor șabloane și tendințe
  - se pot detecta ușor outliers și grupări neobișnuite
- Altă utilizare: reprezentare a datelor obținute după analiză și confruntarea cu cunoștințele unor experți umani sau se pot elimina pattern-urile neinteresante

Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Vizualizarea datelor

Vizualizare - exemplu

Exemplu: date reprezentând temperatura la suprafața apei în Iulie 1982 = zeci de mii de valori.

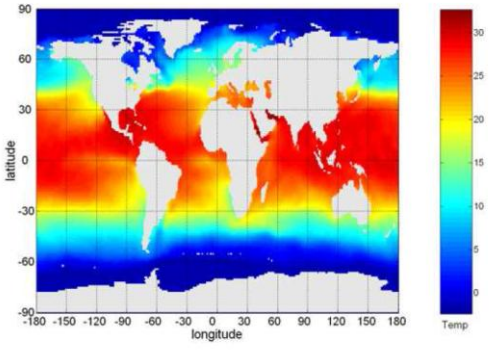
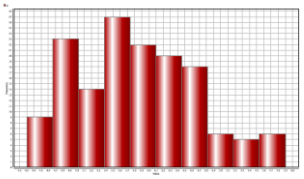


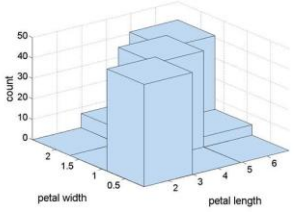
Figure: Rezultat ușor de înțeles și recunoscut: cu cât te îndepărtezi de ecuator, cu atât temperatura scade.

Vizualizarea datelor

Ex. histograme

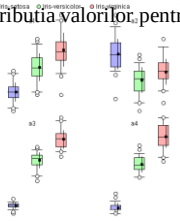


Ex. Histograme bidimensionale

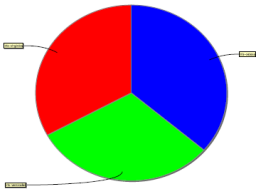


Ex. Matrice boxplots

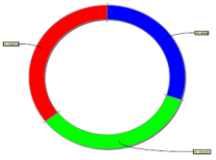
(distributia valorilor pentru un singur atribut numeric



Ex. Pie Chart



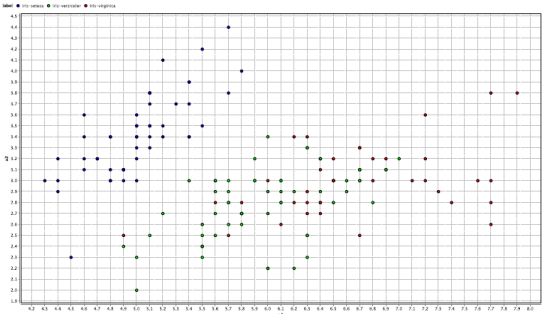
Ex. Ring



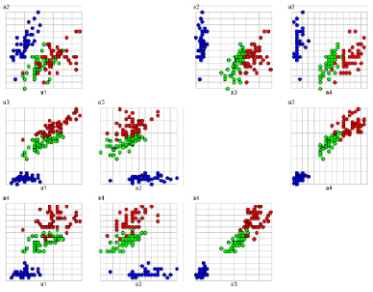
Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Vizualizarea datelor

Ex. Scatter plots

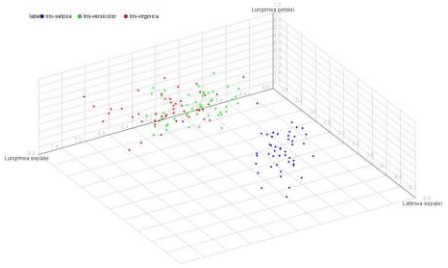


Ex. Matrice Scatter plots



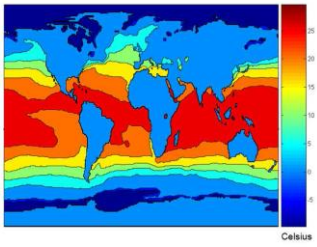
Vizualizarea datelor

Ex. Scatter plots 3D



Ex. Contour plots

Ex. Temperatura medie, decembrie 1998



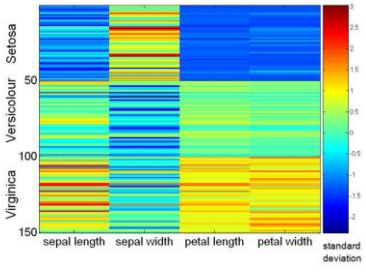
Alte modalitati de vizualizare:

- ❑ Surface plots
- ❑ Vector fields plot
- ❑ Lower dimensional slices
- ❑ Animatii

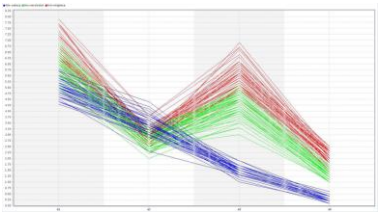
Curs: BAZE DE DATE ÎN INGINERIE ELECTRICĂ

Vizualizarea datelor multidimensionale
Matrici de imagini

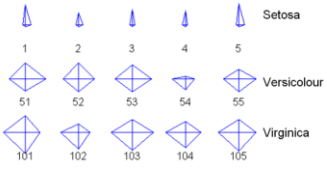
Ex. Matrice de imagini



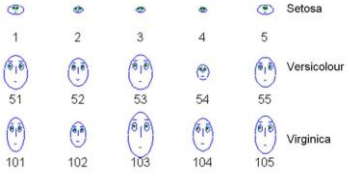
Ex. Coordonate paralele



Ex. Star plots



Ex. fete Chernoff



Limbaje pentru data mining

Languages for analytics / data mining / data science

What programming/statistics languages you used for an analytics / data mining / data science work in 2013? [713 votes total]

	% users in 2013	% users in 2012
R (434 voters in 2013)	60.9%	52.5%
Python (277)	38.8%	36.1%
SQL (261)	36.6%	32.1%
SAS (148)	20.8%	19.7%
Java (118)	16.5%	21.2%
MATLAB (89)	12.5%	13.1%
High-level data mining suite (80)	11.2%	not asked in 2012
Unix shell/awk/sed (79)	11.1%	14.7%
C/C++ (66)	9.3%	14.3%

Instrumente pentru data mining

Data Mining / Analytic Tools Used Poll

Which data mining/analytic tools you used in the past 12 months for a real project (not just evaluation) [912 voters]

RapidMiner (345)	37.8%
R (272)	29.8%
Excel (222)	24.3%
KNIME (175)	19.2%
Your own code (168)	18.4%
Pentaho/Weka (131)	14.3%
SAS (110)	12.0%
MATLAB (84)	9.2%
IBM SPSS Statistics (72)	7.9%
Other free tools (67)	7.3%
IBM SPSS Modeler (former Clementine) (67)	7.3%
Microsoft SQL Server (63)	6.9%
Statsoft Statistica (57)	6.2%
Other commercial tools (56)	6.1%
SAS Enterprise Miner (50)	5.5%
Zementis (34)	3.7%

Instrumente pentru data mining