Prelucrarea datelor cu tehnici de Data Mining

Cuprins

- Ce este data mining
- Etapele procesului de data mining
- Metode si subdomenii in data mining
- Sumar

Definitie ([Liu 11])

- Data mining este cunoscut si sub numele KDD - Knowledge Discovery in Databases (descoperirea de cunostinte in date).
- In mod obisnuit este definit ca procesul de descoperire a unor sabloane/modele (eng. patterns) / cunostinte utile in diverse surse de date: baze de date, colectii de texte, de imagini sau de pagini web, etc.
- Sabloanele trebuie sa fie valide, utile si inteligibile.

Definitie ([Ullman 09, 10])

- Descoperirea de sabloane utile, uneori neasteptate, in date.
- Descoperirea de "modele" pwntru date prin tehnici de tipul:
  - Modelare statistica
  - Invatare automata (Machine learning)
  - Abordari computationale in modelare
  - Sumarizare
  - Extragerea proprietatilor (Feature Extraction)

Definitie ([Wikipedia])

- Data mining (etapa de analiza a procesului KDD - knowledge discovery in databases) este procesul de descoperire de noi sabloane in seturi mari de date prin metode aflate la intersectia dintre inteligenta artificiala, invatarea automata, statistica si baze de date.
- Obiectivul procesului de data mining este acela de a extrage cunostinte dintr-un set de date intr-o forma inteligibila pentru utilizatorii umani.
- Procesul implica baze de date si gestiunea informatiei, preprocesarea datelor, probleme de modelare si inferenta, metrici pentru performanta rezultatelor, probleme de complexitate, post procesare si vizualizare.

Definitie ([Kimball, Ross 02])

- O categorie de cereri executate adesea pe datele atomice pentru a gasi sabloane/modele neasteptate in date.
- Cele mai valoroase rezultate de tip data mining sunt clasificarea, gruparea (clustering) estimarea, predictia si gasirea evenimentelor care apar impreuna.
- Exista multe tipuri de unelte pentru data mining. Cele mai cunoscute sunt arborii de decizie, retelele neuronale, unelte pentru vizualizarea seturilor de date, algoritmi genetici, logica fuzzy si cele folosite in statistica obisnuita.
- In general data mining foloseste date aflate intr-un depozit de date (data warehouse).

Concluzii

- Procesul de data mining converteste datele in cunostinte valoroase care pot fi folosite pentru suportul deciziei.
- Domeniul Data mining consta intr-o colectie de metodologii de analiza a datelor, tehnici si algoritmi pentru descoperirea de noi sabloane.
- Data mining se foloseste cu precadere pentru seturi mari de date.
- Procesul de data mining este automat (nu necesita interventie umana).
- Data mining si Knowledge Discovery in Databases (KDD) sunt considerate de multi autori ca fiind acelasi lucru dar exista autori pentru care care data mining este doar etapa de analiza si de extragere a sabloanelor a procesului de descoperire a cunostintelor in date (KDD), etapa care se desfasoara dupa curatarea si transformarea datelor si inainte de vizualizarea si evaluarea rezultatelor.

Exemple de succes

In cursurile profesorului J. Ullman de la Stanford sunt enumerate cateva exemple de succes ale folosirii tehnicilor de data mining (preluare din [Ullman 03]):

- Arbori de decizie construiti din istoria imprumuturilor bancare pentru a genera algoritmi de acordare a imprumuturilor.
- Sabloane ale comportamentului calatorilor pentru a optimiza vanzarea cu pret redus a locurilor la avion sau a camerelor de hotel.
- “Scutece si bere" S-a observat ca cei care cumpara scutece sunt mai inclinati decat ceilalti sa cumpere bere. Datorita acestui rezultat, cele doua produse se pot plasa intr-un supermarket unul in apropierea celuilalt, astfel incat multi cumparatori vor circula intre cele doua raioane. Plasarea cipsurile pe traseu a dus la cresterea vanzarilor pentru toate cele 3 produse.

Exemple de succes

- Skycat si Sloan Sky Survey: gruparea obiectelor ceresti dupa nivelele radiatiei electromagnetice intr-un numar de benzi a permis identificarea galaxiilor, stelelor apropiate si a altor categorii de obiecte celeste.
- Compararea genotipului persoanelor avand sau neavand o anumita problema de sanatate a permis descoperirea unor gene care sunt asociate cu respectiva problema (de exemplu diabet). In acest fel se pot lua masuri de prevenire in cazul persoanelor care prezinta respectivele caracteristici inainte de declansarea bolii.

Ce nu este Data Mining:

- Cautarea unei persoane in baza de date a unei organizatii.
- Calcularea de valori minime, maxime, medii, sume sau numararea valorilor aflate in tabelele unei baze de date (operatii pur statistice).
- Folosirea unui motor de cautare pentru a gasi referintele asociate numelui tau.

Software pentru DM

In lucrarea ([Mikut, Reischl 11]) programele software pentru data mining sunt clasificate in 9 categorii:

- **Suite pentru data mining** (Data mining suites -DMS) sunt pachete dedicate aplicatiilor de DM. Includ numeroase metode / algoritmi. Exemple:
  - Comerciale: IBM SPSS Modeler, SAS Enterprise Miner, DataEngine, GhostMiner, Knowledge Studio, NAG Data Mining Components, STATISTICA
  - Open source: RapidMiner
- **Pachete de Business intelligence** (Business intelligence packages - BIs) includ functionalitati de baza pentru DM - de exemplu metode statistice pentru aplicatii de business. Exemple:
  - Comerciale: IBM Cognos 8 BI, Oracle DataMining, SAPNetweaver Business Warehouse, Teradata Database, DB2 Data Warehouse from IBM
  - Open source: Pentaho

Software pentru DM

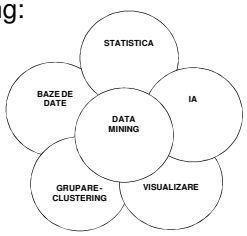
- **Pachete matematice** (Mathematical packages - MATs) contin o multime mare si extensibila de algoritmi precum si rutine de vizualizare. Exemple:
  - Comerciale: MATLAB, R-PLUS
  - Open source: R, Kepler
- **Pachete integratoare** (Integration packages - INTs) sunt colectii extensibile de algoritmi de tip open-source. Exemple:
  - Stand-alone software (KNIME, versiunea GUI pentru WEKA, KEEL, TANAGRA)
  - Extensii pentru uneltele de tip MAT descrise mai sus.
- **Extensii** (EXT) sunt accesorii (add-ons) pentru alte unelte ca Excel, Matlab, R, cu o functionalitate limitata dar utila in diverse cazuri. Exemple:
  - Retele neuronale pentru Excel (Forecaster XL, XLMiner)
  - MATLAB (Matlab Neural Networks Toolbox).

Software pentru DM

- **Biblioteci de DM** (Data mining libraries - LIBs) contin functii care pot fi folosite in alte unelte software. Exemple: Neurofusion for C++, WEKA, MLC++, JAVA Data Mining Package, LibSVM
- **Specialitati** (Specialties - SPECs) sunt similare cu DMS dar implementeaza doar o categorie/familie de metode. Exemple: CART, Bayesia Lab, C5.0, WizRule, Rule Discovery System, MagnumOpus, JavaNNS, Neuroshell, NeuralWorks Predict, RapAnalyst.
- **Cercetare** (Research - RES) sunt primele implementari ale unor noi algoritmi, avand de obicei un suport grafic restrans si fara foarte multe facilitati in utilizare. Sunt de obicei open source. WEKA si RapidMiner au debutat in aceasta categorie.
- **Solutii** (Solutions - SOLs) descriu unelte care sunt adaptate la un domeniu ingust de aplicatii. Exemple: pentru text mining: GATE; pentru prelucrare de imagini: ITK, ImageJ; cercetarea farmaceutica: Molegro Data Modeler

Comunitati implicate

Cele mai importante comunitati implicare in data mining:



Cuprins

- Ce este data mining
- Etapele procesului de data mining
- Metode si subdomenii in data mining
- Sumar

Etapele procesului de DM (1)

1. **Colectarea datelor**: Datele sunt culese din baze de date existente sau prin parcurgerea paginilor web (Web crawling).
2. **Preprocesarea datelor**, presupune mai multe activitati desfasurate in scopul pregatirii datelor pentru aplicarea algoritmilor specifici

Preprocesarea datelor

- Curatarea datelor (Data cleaning):
  - inlocuirea sau inlaturarea valorilor lipsa,
  - netezirea datelor care contin zgomot,
  - identificarea si eventual inlaturarea punctelor singulare (outliers),
  - inlaturarea inconsistentelor.

Preprocesarea datelor

- Integrarea datelor (Data integration):
  - integrarea intr-un corp unic a datelor provenite din mai multe surse,
  - Conversii de tip si de structura
  - eliminarea duplicatelor si
  - tratarea inconsistentelor datorate integrarii

Preprocesarea datelor

- Transformarea datelor (Data transformation):
  - normalizarea (sau standardizarea) datelor,
  - sumarizari,
  - generalizari, c
  - onstructia de noi atribute, etc.

Preprocesarea datelor

- Reducerea datelor (Data reduction): nu toate atributele existente sunt necesare pentru un anumit proces de data mining.
- Prin reducere se opresc doar acele atribute care sunt necesare diminuand astfel volumul de date prelucrate (si implicit timpul de rulare al algoritmului).

Preprocesarea datelor

- Distretizare: Anumiti algoritmi lucreaza doar pe date discrele. De aceea pentru atributele avand valori continue trebuie efectuata discretizarea constand in inlocuirea acestor valori cu altele dintr-o multime de valori discrete.
- De exemplu varsta se poate inlocui cu un atribut avand doar trei valori: Tanar, Varsta-mijlocie si Batran.

Etapele procesului de DM (2)

3. **Extragerea sabloanelor si descoperirea de cunostinte**. Aceasta este etapa in care sunt efectiv utilizati algoritmii de DM pentru obtinerea rezultatului. Unii autori reduc domeniul Data Mining la aceasta etapa, intregul proces fiind KDD.
4. **Vizualizarea:** deoarece data mining extrage proprietati/informatii ascunse din date este necesara o vizualizare a rezultatelor pentru o mai buna intelegere a lor si pentru a le evalua. Vizualizarea este uneori necesara si pentru datele de intrare.

Etapele procesului de DM (3)

3. **Evaluarea rezultatului:** nu tot ce iese dintr-un proces de data mining este valoros.

- Unele rezultate sunt adevaruri pur statistice care se puteau deduce si fara aplicarea algoritmilor iar altele nu prezinta interes pentru utilizator.
- De aceea este necesara evaluarea rezultatelor de catre experti pentru a separa cunostintele valoroase de celelalte lucruri obtinute in urma rularii algoritmului.

Principiul lui Bonferroni (1)

O informatie adevarata descoperita in urma procesului de data mining poate fi un adevar pur statistic. Exemplu (din [Ullman 03]):

- In 1950 David Rhine, un parapsiholog de la universitatea Duke a testat studentii pentru a afla daca au sau nu perceptie extrasenzoriala (ESP).
- Pentru asta le-a cerut sa ghiceasca culoarea a 10 carti de joc succesive - rosu sau negru. Rezultatul a fost ca 1/1000 din participanti au ghicit toate cele 10 carti - in consecinta el i-a declarat ca avand ESP.
- Re-testarea efectuata doar asupra acestora nu a mai avut aceleasi rezultate, el considerand ca in momentul in care au aflat ca au ESP au pierdut aceasta capacitate.
- David Rhine nu a realizat ca statistica spune ca probabilitatea de a ghici 10 carti succesive este de $1/1024 = 1/2^{10}$, deoarece probabilitatea de a ghici o carte este 1/2 (rosu sau negru).

Principiul lui Bonferroni (1)

- Astfel de rezultate pot fi regasite in iesirea unui algoritm de data mining si trebuiesc recunoscute ca adevaruri statistice si nu ca rezultate reale ale procesului de data mining.
- Astfel de rezultate sunt obiectul principiului lui Bonferroni. Acesta poate fi sintetizat astfel:

Daca sunt prea multe concluzii, unele vor fi adevarate din motive pur statistice

Cuprins

2 tipuri de metode

- **Metode de predictie:** Aceste metode utilizeaza anumite variabile pentru a prezice valoarea altor variabile. Un exemplu din aceasta categorie este clasificarea: bazat pe date cunoscute si deja etichetate (clasificate), algoritmii de clasificare creaza modele care pot fi folosite pentru clasificarea unor date noi, necunoscute.
- **Metode descriptive:** Algoritmii din aceasta categorie gasesc sabloane / modele care descriu structura interna a unui set de date. De exemplu algoritmii de grupare (clustering) gasesc grupuri de obiecte similare in setul de date (grupuri numite clustere) si de asemenea detecteaza posibile obiecte izolate, departate de oricare dintre clustere, numite si puncte singulare (outliers).

Algoritmi

- **Algoritmi de predictie:**
  - Clasificare
  - Regresie
  - Detectia deviatiei
- **Algoritmi de descriere:**
  - Grupare - Clustering
  - Gasirea de reguli de asociere
  - Descoperirea de sabloane secventiale

Clasificare

Input:

- Un set avand un numar k de clase $C = \{c_1, c_2, \ldots, c_k\}$
- Un set de n articole etichetate $D = \{(d_1, c_{i1}), (d_2, c_{i2}), \ldots, (d_n c_{in})\}$. Articolele sunt $d_1, \ldots, d_n$, fiecare articol d_j fiind etichetat cu clasa $c_{ij} \in C$. D este numit si setul (multimea) de **antrenare (training set).**
- Pentru calibrarea unor algoritmi este necesar si un set **de validare (validation set)**. Acesta contine de asemenea articole etichetate, neincluse in setul de antrenare

Output:

- Un model sau metoda de a clasifica noi articole (clasificator). Setul continand noile articole se numeste set **de test (test set)**

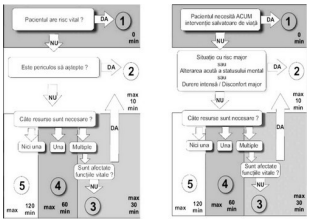
Exemplu

- Sa consideram ca articolele sunt pacienti ai unui serviciu de urgente.
- Exista 5 clase, C0, C10, C30, C60 si C120, unde eticheta Ck inseamna ca pacientul poate fi lasat sa astepte maxim k minute.
- Vom reprezenta datele setului de antrenare sub forma tabelara.
- Iesirea unui algoritm de creare a unui clasificator poate fi de exemplu un arbore de decizie sau un set ordonat de reguli.
- Modelul obtinut poate fi utilizat apoi pentru a clasifica noi pacienti asignandu-le o eticheta din multimea celor 5 de mai sus.

Multimea de antrenare UPU

Name (or ID)	Vital risk?	Danger if waits?	0 resource needed	1 resource needed	>1 resource needed	>1 resource needed and vital function s affected	Waiting time (class label)
John	Yes	Yes	No	Yes	No	No	C0
Maria	No	Yes	No	No	Yes	No	C10
Nadia	Yes	Yes	Yes	No	No	No	C0
Omar	No	No	No	No	Yes	Yes	C30
Kiril	No	No	No	Yes	No	Yes	C60
Denis	No	No	No	No	Yes	No	C10
Jean	No	No	Yes	Yes	No	No	C120
Patricia	Yes	Yes	No	No	Yes	Yes	C60

Rezultat: arbori de decizie



- Pentru un nou pacient:

Felix	Yes	Yes	No	No	No	Yes	?????

Arborele de decizie produce clasa C0

Regresia (1)

- Regresia provine din statistica.
- Inseamna: prezicerea valorii unei anumite variabile continue pe baza valorilor altor variabile, considerand un model de dependenta liniara sau neliniara ([Tan, Steinbach, Kumar 06]).
- Utilizata in predictie si prognoza; este folosita si in domeniul invatarii automate.
- Analizele bazate pe regresie sunt folosite pentru intelegerea relatiilor dintre variabile dependente si independente.

Regresia (2)

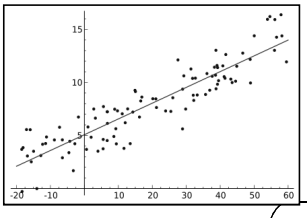
Exista multe tipuri de regresie, de exemplu:

- Regresie liniara
- Regresie liniara simpla
- Regresie logistica
- Regresie neliniara
- Regresie nonparametrica
- Regresie robusta

Exemplu

Exemplu de regresie liniara

(sursa: http://en.wikipedia.org/wiki/File:Linear_regression.svg)



Detectia deviatiei

- Detectia deviatiei (sau a anomaliilor) presupune descoperirea de deviatii semnificative de la comportamentul normal. Punctele singulare (outliers) sunt o categorie semnificativa de astfel de date anormale.
- Detectia deviatiei poate fi utilizata in multe situatii:
  - In faza de rulare a algoritmului de data mining, datele anormale putand avea un efect puternic asupra algoritmului
  - Audit: astfel de informatii pot arata existenta unor probleme sau practici gresite
  - Detectia fraudelor: cererile frauduloase contin deseori informatii inconsistente
  - Detectia intruziunilor intr-o retea de calculatoare poate fi facuta si pe baza unor date anormale
  - Curatarea datelor (data cleaning): astfel de informatii anormale pot reprezenta date eronate care trebuiesc corectate

Metode de detectie a deviatiei

- Tehnici bazate pe distante (ex. : k-nearest neighbor).
- Folosirea "One Class Support Vector Machines" (Ca un SVM clasic dar toate exemplele de antrenare sunt din aceeasi clasa si doar originea reprezinta cea de-a doua clasa).
- Metode predictive (arbori de decizie, retele neuronale).
- Detectia punctelor singulare folosind clustering.
- Inregistrari care nu respecta regulile de asociere
- Analize Hotspot (clustere avand o valoare ridicata a anumitor parametri; de exemplu politia foloseste astfel de analize pentru analiza zonelor unde criminalitatea are valori ridicate)

Algoritmi

- **Algoritmi de predictie:**
  - Clasificare
  - Regresie
  - Detectia deviatiei
- **Algoritmi de descriere:**
  - Grupare - Clustering
  - Gasirea de reguli de asociere
  - Descoperirea de sabloane secventiale

Clustering

Input:

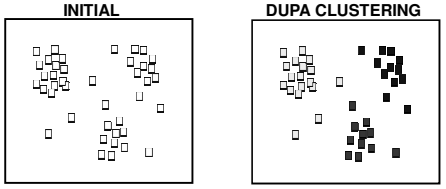
- Un set de obiecte D = {d_1, d_2, …, d_n} (numite uzual puncte). Obiectele nu sunt etichetate si nu exista definit vreun set de clase.
- O functie de distanta (masura a disimilaritatii) care poate fi utilizata pentru a calcula distanta dintre oricare doua puncte. O distanta mica inseamna "aproape" iar una mare "departe".
- Unii algoritmi necesita introducerea unei valori pentru numarul de clustere de obtinut.

Output:

- Un set de grupuri de obiecte/puncte, numite clustere unde punctele din acelasi cluster sunt 'apropiate' iar cele din clustere diferite 'departate' unele de altele, considerand functia de distanta existenta.

Exemplu

- Daca avem un set de 2 puncte intr-un spatiu 2D, sa se gaseasca grupurile/clusterele naturale formate de ele



Sursa: http://en.wikipedia.org/wiki/File:Cluster-1.svg, http://en.wikipedia.org/wiki/File:Cluster-2.svg

Reguli de asociere

Input:

- Un set de **m** articole I = {i_1, i_2, …, i_m}.
- Un set de **n** tranzactii T = { t_1, t_2, …, t_n}, fiecare tranzactie continand un subset al lui I. Deci daca $t_k \in T$ atunci t_k = {i_{k1}, i_{k2}, …, i_{kj}} unde j depinde de k (tranzactiile au lungimi diferite).
- Un prag s numit prag de suport, dat fie in procente fie in valoare absoluta. Daca o multime de articole (itemset) $X \in I$ este inclusa in w tranzactii atunci w este suportul lui X. Daca w >= s atunci X este numita multime frecventa de articole.
- Un al doilea prag c pentru increderea regulilor obtinute.

Output:

- Multimile frecvente de articole din T, avand suportul >= s
- Multimea de reguli derivate din T avand suportul >= s si increderea >= c

Reguli de asociere

- O **regula** este o constructie de tipul $X \rightarrow Y$ unde X si Y sunt multimi de articole (itemsets).
- **Suportul** unei reguli $X \rightarrow Y$ este numarul de aparitii ale $X \cup Y$ in T (egal cu suportul acestei reuniuni ca itemset):

$$\text{support}(X \rightarrow Y) = \text{support}(X \cup Y)$$

- Increderea unei reguli este proportia tranzactiilor continandu-l pe Y in multimea trabzactiilor continandu-l pe X:

$$\text{confidence}(X \rightarrow Y) = \text{support}(X \cup Y) \,/\, \text{support}(X).$$

- Acceptam o regula ca valida doar daca suportul si increderea ei sunt mai mari sau egale cu pragurile date.

Exemplu

- Sa consideram urmatorul set de tranzactii:

Transaction ID	Items
1	Bread, Milk, Butter, Orange Juice, Onion, Beer
2	Bread, Milk, Butter, Onion, Garlic, Beer, Orange Juice, Shirt, Pen, Ink, Baby diapers
3	Milk, Butter, Onion, Garlic, Beer
4	Orange Juice, Shirt, Shoes, Bread, Milk
5	Butter, Onion, Garlic, Beer, Orange Juice

- Daca s = 60% atunci {Bread, Milk, Orange Juice} sau {Onion, Garlick, Beer} sunt multimi frecvente de articole.
- De asemenea, daca s = 60%, c=70% atunci regula {Onion, Beer} → {Garlic} e valida avand suport de 60% si incredere de 75%.

Descoperire sabloane secventiale

Input:

- O multime de **m** articole I = $\{i_1, i_2, \ldots, i_m\}$. O secventa este o lista ordonata de multimi de articole din I.
- O multime de secvente **S** (numita si 'sequence database').
- O functie booleana care poate testa daca o secventa S_1 este inclusa (este o subsecventa) in secventa S_2. In acest caz S_2 este numita o supersecventa a lui S_1.
- Un prag **s** (procent sau valoare absoluta) necesar pentru a gasi secvente frecvente.

Output:

- Multimea de secvente frecvente, i.e. multinea de secvente incluse in cel putin **s** secvente din S.
- Uneori se poate obtine si un set de reguli, fiecare regula fiind de forma $S_1 \rightarrow S_2$ unde S_1 si S_2 sunt secvente.

Exemplu

- Intr-o librarie putem gasi secvente ca:

  $\{Book\_on\_C, Book\_on\_C++, Book\_on\_Perl\}$

- Din aceasta secventa se poate deriva o regula de tipul: dupa cumpararea unor carti de C si C++, clientul cumpara carti de Perl:

  $Book\_on\_C, Book\_on\_C++ \rightarrow Book\_on\_Perl$

Sumar

In acest curs am prezentat:

- O lista de definitii alternative pentru Data Mining si cateva exemple pentru a intelege ce este Data Mining si ce nu este Data Mining.
- O discutie despre comunitatile implicate in Data Mining si despre faptul ca Data Mining este de fapt o reuniune heterogena de subdomenii.
- Pasii procesului de Data Mining process de la colectarea datelor aflate in locatii diferite (depozite de date, arhive sau sisteme operationale) pana la pasul final de evaluare
- O scurta descriere a subdomeniilor principale ale Data Mining cu exemple pentru fiecare dintre ele.

Saptamana viitoare: Multimi frecvente de articole si reguli de asociere

Peste doua saptamani: Lucrarea de la mijlocul semestrului

Bibliografie si referinte

- [Liu 11] Bing Liu, 2011. Web Data Mining, Exploring Hyperlinks, Contents, and Usage Data, Second Edition, Springer, 1-13.
- [Tan, Steinbach, Kumar 06] Pang-Ning Tan, Michael Steinbach, Vipin Kumar, 2006. Introduction to Data Mining, Adisson-Wesley, 1-16.
- [Kimbal, Ross 02] Ralph Kimball, Margy Ross, 2002. The Data Warehouse Toolkit, Second Edition, John Wiley and Sons, 1-16, 396
- [Mikut, Reischl 11] Ralf Mikut and Markus Reischl, Data mining tools, 2011, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, Volume 1, Issue 5, http://onlinelibrary.wiley.com/doi/10.1002/widm.24/pdf
- [Ullman] Jeffrey Ullman, Data Mining Lecture Notes, 2003-2009, web page: http://infolab.stanford.edu/~ullman/mining/mining.html
- [Wikipedia] Wikipedia, the free encyclopedia, en.wikipedia.org