ASUPRA TERMENULUI DE MINERIT DE DATE (*DATA MINING*)

Angela Ioniță

Institutul de Cercetări pentru Inteligență Artificială, Academia Română

Rezumat: **Mineritul de date** (data mining) s-a dezvoltat ca o consecință a disponibilizării marilor rezervoare de date. Colectarea datelor în diverse formate de digitizare a început în anii '60 permițând o analiză retrospectivă a datelor prin intermediul calculatorului. Bazele de date relaționale au apărut în anii '80 împreună cu Structured Query Language (SQL) permițând analizarea dinamică la cerere a datelor. Anii '90 sunt caracterizați de o explozie a datelor. Pentru stocarea lor au început să se folosească depozitele de date (data warehouses). Mineritul de date a apărut ca răspuns la provocările cu care s-a confruntat comunitatea specialiștilor în baze de date, care se ocupau cu cantități masive de date, aplicarea analizei statistice și aplicare tehnicilor de căutare, specifice inteligenței artificiale asupra datelor. **Mineritul de date** este aplicat într-o varietate de domenii, începând cu managementul de investiții până la astronomie. Importanța și potențialul de aplicare al mineritului de date a fost recunoscut în marketing, domeniul bancar, asigurarea sănătății, telecomunicații ș.a. pentru aplicații cum ar fi analiza coșului de piață, pentru promovarea eficienței, analiza vulnerabilității clienților, managementul relațiilor cu clienții, crearea de portofoliu, detectarea fraudei în telefonia celulară etc. În fiecare dintre aceste aplicații este necesară executarea mai multor operații de minerit de date decât în domeniile depozitării de date (data warehousing) și sistemelor suport pentru decizie. Întrucât până la această dată încă nu există consens asupra traducerii și utilizării termenului de **minerit de date** (data mining), acest articol și-a propus discutarea mai multor definiții mai mult sau mai puțin acceptate în diferite comunități de specialiști și a contextelor de utilizare. Datorită evoluției rapide a accesării datelor online datorată dezvoltării Internet-ului, s-a creat o imensă cerere de metodologii de descoperire de cunoștințe. În consecință, terminologia a evoluat și ea, **mineritul de date** căpătând diferite înțelesuri, așa cum este prezentat în prima secțiune a acestui articol. Cea de a doua secțiune face o prezentare a mineritului de date ca etapă în procesul de extragere de cunoștințe și este urmată de o foarte scurtă prezentare a celor mai utilizați algoritmi. În secțiunea a patra, sunt prezentate câteva clase de probleme cărora li se adresează **mineritul de date**. Secțiunea a cincea se referă la tehnologiile de minerit de date. Ultima secțiune este dedicată concluziilor, punctând asupra înțelesului actual al termenului, tendințelor de standardizare și asupra unor aspecte caracteristice. Fără a avea pretenția de exhaustivitate, acest articol are intenția de a atrage atenția asupra unui domeniu nou, în plină dezvoltare, al științei calculatoarelor, care va furniza un nivel nou și eficient de informații și de descoperire de cunoștințe de care vor beneficia toți utilizatorii din domeniul memorării computerizate de date.

Cuvinte cheie: minerit de date (data mining), depozitare de date (data warehousing), descoperirea de cunoștințe (knowledge discovery), baze de date, reguli de asociere, clusterizare, algoritmi de clasificare, arbori de decizie, rețele neuronale, algoritmi genetici.

1. Introducere

În 1995, Gartner Group Advanced Technology Research Note a listat **mineritul de date** și inteligența artificială pe primul loc între cele cinci zone tehnologice cheie care „vor avea în mod clar impact peste un spectru larg de domenii industriale în următorii 3 până la 5 ani."[1] Această predicție s-a dovedit a fi adevărată pe măsură ce o serie de produse de minerit de date au fost introduse pe piață și multe domenii economice au început să le folosească cu regularitate, beneficiind din utilizarea acestor produse.

Mineritul de date **(data mining = DM), cunoscut și ca *descoperire de cunoștințe în baze de date* (knowledge-discovery in databases = KDD),** este practica de căutare automată de șabloane în depozite mari de date (Wikipedia article „*Data mining*"[2]). În vederea realizării acestui scop, mineritul de date utilizează statistica și recunoașterea de forme (pattern recognition).

Mineritul de date a fost definit ca fiind „*extragerea de informații netriviale, necunoscute anterior și potențial utile din date*" [9], dar și ca fiind „*știința extragerii de informații utile din masive de date sau baze de date*" [11].

Mineritul de date este un termen acoperitor și este folosit într-un spectru larg de contexte cu înțelesuri diferite.

Folosit în contextual tehnic al depozitelor de date (data warehousing) și analizei, mineritul de date este neutru. Totuși, uneori termenul a fost utilizat în sens peiorativ, impunând șabloane (și, în particular, relații cauzale) pe date, acolo unde ele nu existau. Această impunere de corelații nerelevante, care induc erori sau corelații triviale de atribut, este mult criticatul termen din statistică „*dragare de date*" (data dredging). În sens restrâns, „*dragarea de date*" implică scanarea datelor pentru orice relații și, când se găsește ceva, se dă o explicație interesantă. Problema constă în faptul că, în mod invariabil, se întâmplă ca în masivele de date să existe relații particulare interesante. Un alt pericol constă în descoperirea de corelații care nu există în mod real. Analiștii de investiții sunt cei care sunt mai vulnerabili în această zonă.

În **mineritul de date** s-a depus mult efort în dezvoltarea unui model de granularitate fină și cât se poate de detaliat pentru masivele de date. În „*Data Mining For Very Busy People*" [14], cercetătorii de la West Virginia University și

---

[1] www3.shore.net/~kht/text/dmwhite.dmwhite.htm

[2] http://en.wikipedia.org/wiki/Data mining

University of British Columbia au discutat o metodă care implică găsirea de diferențe minimale între elementele unei mulțimi de date, cu scopul de a dezvolta modele simple care reprezintă date relevante.

Există o oarecare reținere în ceea ce privește aplicarea **mineritului de date**. De exemplu, dacă un angajator are acces la înregistrările medicale, legate de persoanele care au diabet sau au suferit un atac de cord, atunci se vor putea micșora sau tăia cheltuielile de asigurare, dar se vor naște probleme de etică și probleme legale. **Mineritul de date** în mulțimile de date guvernamentale sau comerciale în scopul asigurării securității statului sau aplicării legii implică, de asemenea, multă prudență [19].

Există, însă, multe utilizări îndreptățite ale **mineritului de date**. De exemplu, o bază de date a medicamentelor prescrise unei categorii de pacienți poate fi utilizată pentru a găsi combinații de medicamente care au anumite reacții adverse. Dacă o anume combinație apare doar o dată la 1000 de pacienți, cazul poate să nu fie semnificativ. Un proiect care include farmaciile poate să reducă numărul de medicamente cu reacții adverse și poate să salveze vieți. Din nefericire, posibilitatea utilizării abuzive a acestui tip de baze de date există.

În esență, **mineritul de date** dă informații care altfel nu ar putea fi disponibile. Pentru a putea fi utile, ele trebuie să fie interpretate corect. Atunci când datele colectate implică și persoane, apar mai multe probleme legate de confidențialitate, intimitate, legalitate și etică.

Mineritul de date constă dintr-o mulțime de tehnici în continuă dezvoltare care pot fi utilizate pentru a extrage informații valoroase și cunoștințe, din volume masive de date. Până la un moment dat, cercetările din **mineritul de date** și instrumentele au pus accentul mai mult pe aplicațiile comerciale. Puține cercetări s-au desfășurat punând accentul pe datele științifice și datele satelitare. Deși în mai multe conferințe dedicate diverselor aspecte implicate de mineritul de date s-a discutat și despre mineritul datelor științifice, nu a existat un schimb de idei concertat pe mineritul de date științifice între oamenii de știință și comunitatea de specialiști în mineritul de date. (www.cs.uah.edu/~thinke/NASA_Mining/DMFinalReport.pdf).

Mineritul de date este procesul prin care informațiile și cunoștințele sunt extrase din volumele mari de date folosind tehnici care sunt mai mult decât o simplă căutare în date[3]. [15].

Mineritul de date este o etapă în procesul de descoperire de cunoștințe, care constă din aplicarea analizei de date și a algoritmilor de descoperire care, sub limite rezonabile ale eficienței de calcul, produce o enumerare particulară de șabloane (sau modele) de date. De menționat că spațiul metodelor, numărul efectiv de variabile luate în considerare poate fi redus sau se pot găsi reprezentări invariante pentru date.

„**Mineritul de date** este procesul de identificare de cunoștințe valide, noi, potențial utile și, în final, inteligibile din baze de date care sunt folosite la luarea deciziilor hotărâtoare în domeniul afacerilor"[4] [18].

În conformitate cu Carta Albă a soluțiilor de management de date de la IBM [12], **mineritul de date** este procesul de extragere de informații valide, necunoscute anterior și, în final, inteligibile din baze mari de date, informații folosite în luarea deciziilor hotărâtoare în domeniul afacerilor. Extragerea informațiilor poate fi utilizată la formarea de modele de predicție sau de clasificare, pentru a identifica relații între înregistrările din bazele de date sau pentru a furniza un sumar al bazelor de date care sunt minerite. **Mineritul de date** constă dintr-un număr de operații, fiecare fiind suportată de o varietate de tehnici cum sunt reguli de inducție, rețele neuronale, clusterizare conceptuală, descoperire asociativă etc. În multe domenii din lumea reală cum ar fi analiza de marketing, analiza financiară, detectarea fraudei ș.a. informațiile extrase necesită utilizarea cooperativă a mai multor operații și tehnici de minerit de date.

„**Mineritul de date** este un process de descoperire de relații, șabloane și cunoștințe din date"[5]. [21].

---

[3] *"Data mining is the process by which information and knowledge are extracted from a potentially large volume of data using techniques that go beyond a simple search through the data".*

[4] *"Data Mining is the process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge from database that is used to make crucial business decisions".*

[5] *"Data mining is a process of discovery of relationships, patterns and knowledge from data".*

2. Mineritul de date şi descoperirea de cunoştinţe în baze de date

2.1. Date, informaţii, cunoştinţe

2.1.1. Date

Datele sunt fapte oarecare, numere sau texte care pot fi procesate de un calculator. Organizaţiile de astăzi sunt într-o continuă acumulare de cantităţi mari de date, în diferite formate şi în diferite baze de date. Acestea includ:

- date operaţionale sau tranzacţionale cum ar fi vânzări, costuri, inventare, plăţi, contabilitate;
- date nonoperaţionale cum ar fi vânzări industriale, predicţii, date macroeconomice;
- metadate – date despre date cum ar fi proiectarea logică a bazei de date sau definiţii ale dicţionarului de date.

2.1.2. Informaţii

Şabloanele, asocierile sau relaţiile între toate aceste *date* pot furniza *informaţii*. De exemplu, analiza în punctul de vânzare cu amănuntul a datelor referitoare la tranzacţia vânzare poate să scoată la iveală informaţii despre ce produse sunt vândute şi unde.

2.1.3. Cunoştinţe

Informaţiile pot fi convertite în *cunoştinţe* despre şabloane istorice şi tendinţe de viitor. De exemplu, sintezele informative despre vânzările din supermarket-uri de vânzare cu amănuntul pot fi analizate din punctual de vedere al eforturilor promoţionale pentru a furniza cunoştinţe despre comportamentul consumatorilor. Deci, un producător sau un distribuitor poate determina care articole sunt cele mai susceptibile la eforturile promoţionale.

2.1.4. Depozite de date

Progresele dramatice în capturarea de date, puterea de procesare, transmiterea de date şi capabilităţile de stocare au permis organizaţiilor să-şi integreze diferitele lor baze de date în *depozite de date*. Depozitarea de date este definită ca un proces de management centralizat de date şi de regăsire. Depozitarea de date, ca şi mineritul de date, este un termen relativ nou, deşi conceptual a apărut de mai mulţi ani. Depozitarea de date reprezintă o viziune ideală asupra întreţinerii unui depozit central al tuturor datelor organizaţionale. Centralizarea datelor este necesară pentru maximizarea accesului utilizatorilor şi pentru analiză. Progresele tehnologice au făcut din această viziune o realitate pentru multe companii. Progresele făcute în dezvoltarea software-ului de analiză permit utilizatorilor să acceseze liber datele. Software-ul de analiză de date este cel care sprijină mineritul de date.

În general, mineritul de date este definit şi ca fiind *procesul de analizare de date din diferite perspective şi de sumarizare a lor în informaţie utilă* – informaţie care poate fi utilizată pentru creşterea venitului, micşorarea costurilor etc. Software-ul de minerit de date este format dintr-un număr de instrumente analitice pentru analizarea datelor, care permit utilizatorilor să analizeze datele pe diferite dimensiuni sau din diferite unghiuri, să le clasifice şi să sumarizeze relaţiile identificate. Din punct de vedere tehnic, mineritul de date este *un proces de găsire de corelaţii sau şabloane între numeroase câmpuri din baze de date relaţionale mari* (http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm)

„**Mineritul de date** *este explorarea şi analizarea, prin metode automate sau semiautomate, a unor cantităţi masive de date pentru a descoperi şabloane şi reguli semnificative*" [19].

Deşi există mai multe definiţii acceptate pentru **mineritul de date**, cea de mai sus înglobează faptul că specialiştii în mineritul de date caută şabloane semnificative în cantităţi mari de date. Scopul implicit al unui astfel de efort este acela de a utiliza aceste şabloane semnificative în vederea îmbunătăţirii practicilor din domeniul afacerilor incluzând marketing, vânzări şi management personalizat. Din punct de vedere istoric găsirea de şabloane utile în date a fost referită, pe lângă minerit de date, ca fiind şi descoperirea de cunoştinţe, descoperirea de informaţii, recoltarea de informaţii, arheologie de date şi procesarea de şabloane de date. În ultimii ani, s-a stabilit că termenul care descrie aceste activităţi este **mineritul de date** [6]. Statisticienii au utilizat termenul de minerit de date pentru a referi şabloanele de date care sunt descoperite prin analiză de regresie şi alte tehnici statistice.

Pe măsură ce **mineritul de date** s-a maturizat, s-a acceptat faptul că **mineritul de date** este o fază în cadrul ciclului de viaţă al *descoperirii cunoştinţelor în baze de date* (Knowledge Discovery in Databases = KDD). Termenul de *descoperire de cunoştinţe în baze de date* a fost inventat în 1989 pentru a face referire în sens larg la găsirea de cunoştinţe în depozitele de date [7]. Domeniul *descoperirii cunoştinţelor în baze de date* este orientat, în special, pe activităţile care conduc la analiza de date, incluzând evaluarea şi extinderea rezultatelor. Descoperirea de cunoştinţe în baze de date cuprinde următoarele activităţi (figura 1):

1. **Selecţia de date** – Scopul acestei faze este acela de a extrage din masive de date numai datele care sunt relevante pentru analiză în mineritul de date. Această extragere de date ajută la canalizarea şi creşterea vitezei procesului.
2. **Procesarea de date** – Această fază a descoperirii cunoştinţelor în baze de date se ocupă de curăţirea datelor şi pregătirea activităţilor care sunt necesare în asigurarea de rezultate corecte. Eliminarea lipsei valorilor în date, asigurarea că valorile codificate au un înţeles uniform şi asigurarea faptului că nu există valori greşite sunt acţiuni tipice care apar în această fază.
3. **Transformarea de date** – Această fază a ciclului de viaţă are ca scop convertirea datelor într-o tabelă bidimensională şi eliminarea câmpurilor nedorite sau înalt corelate astfel încât rezultatele să fie valide.
4. **Mineritul de date** – Scopul fazei de minerit de date este de a analiza datele printr-un set de algoritmi potriviţi pentru a descoperi şabloane şi reguli semnificative şi pentru a produce modele predictive. Acesta este nucleul ciclului descoperirii cunoştinţelor în baze de date.
5. **Interpretarea şi Evaluarea** – Deşi algoritmii de minerit de date au puterea de a produce un număr nelimitat de şabloane ascunse, multe dintre acestea nu pot fi semnificative sau utilizabile. Această fază finală are ca scop selectarea acelor modele care sunt valide şi utile în viitoarele decizii din diferite domenii.

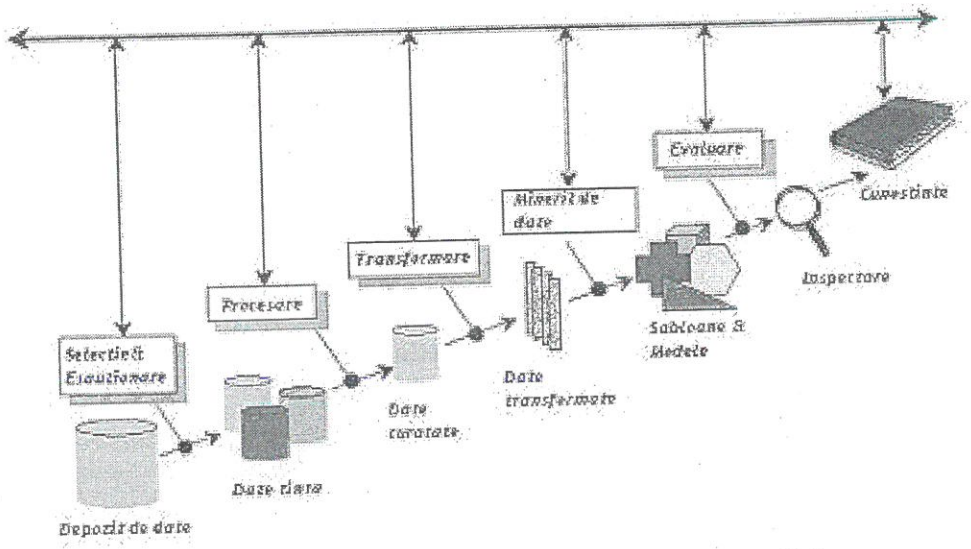

Figura 1: Paradigma clasică a descoperirii de cunoştinţe
(preluată din K. Collier, B. Carey, E. Grusy, C. Marjaniemi, D. Sautter, 1998)

Rezultatul acestui proces sunt cunoştinţele recent achiziţionate care au fost ascunse în date. Aceste cunoştinţe noi pot fi utilizate în viitor în luarea de decizii din diferite domenii.

Acest model de proces poate fi extins astfel:

- **Încadrarea interogărilor** – Una dintre cele mai răspândite dintre percepţiile greşite din **mineritul de date** constă în faptul că se poate ca setul de algoritmi să se aplice orbeşte asupra datelor pentru a găsi toate şabloanele interesante.
- **Mineritul de date** nu este o soluţie universal valabilă. Mai degrabă este un instrument suport pentru decizie care, atunci când este utilizat în conjuncţie cu înţelegerea domeniului de afaceri, poate să furnizeze o metodă valoroasă pentru a spori înţelegerea în domeniu. De aceea, primul pas în ciclul descoperirii de cunoştinţe în baze de date trebuie să discearnă între una sau mai multe interogări pentru a ajuta direct focalizarea în descoperirea de cunoştinţe în baze de date.
- **Rezultate furnizate** – Ciclul de descoperire de cunoştinţe în baze de date se termină cu evaluarea şi validarea rezultatelor analitice. Dificultatea constă în faptul că nu se poate face o recomandare asupra a ceea ce se poate face cu aceste rezultate pentru a sprijini deciziile din domeniul afacerilor.
- **Iteraţia** – Deşi ciclul descoperirii de cunoştinţe în baze de date acceptă revenirea la faze anterioare pentru a îmbunătăţi rezultatele provenite din mineritul de date, experienţa a arătat că iteraţia este mai mult un element integrator în ciclu decât un element implicat de modelul tradiţional. Cu timpul, a fost adoptat un model al procesului de descoperire de cunoştinţe în baze de date, care încorporează faze suplimentare (figura 2). În cadrul acestui model, se poate face un studiu pentru a determina dacă datele vor sprijini obiectivele. Iteraţiile succesive vor servi la rafinarea şi ajustarea datelor şi algoritmilor îmbunătăţi rezultatele.

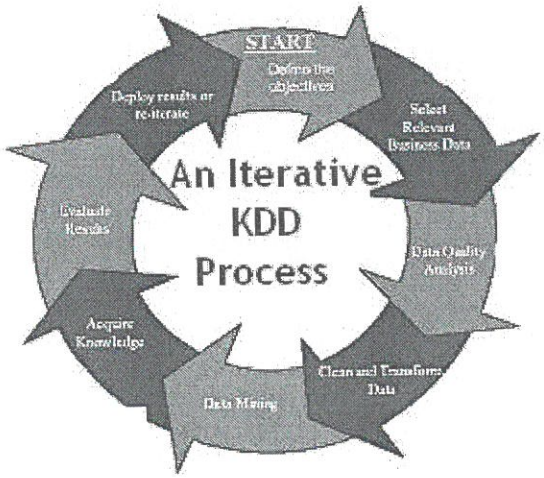

Figura 2: O paradigmă rafinată a descoperirii de cunoştinţe
(preluată din K. Collier, B. Carey, E. Grusy, C. Marjaniemi, D. Sautter, 1998)

3. Algoritmii utilizaţi în mineritul de date

Aşa cum este prezentat în sinteză în [4], principalii algoritmi utilizaţi în mineritul de date sunt:

Reguli de asociere	Identifică relaţii de tip cauză – efect şi atribuie probabilităţi sau factori de încredere pentru a sprijini concluziile. Regulile sunt de forma "if *<condiţie>*, then *<concluzie>*" şi pot fi utilizate pentru a face estimări asupra valorilor necunoscute.
Memory-based Reasoning (MBR) sau Case-based Reasoning (CBR)	Aceşti algoritmi găsesc cele mai vechi analogii ascunse ale unei situaţii prezente pentru a estima o valoare necunoscută sau un rezultat necunoscut.
Analiză de cluster	Separă datele eterogene în subgrupuri omogene şi semiomogene. Bazându-se pe presupunerea că observaţiile tind să fie asemănătoare pe vecinătăţi, clusterizarea îmbunătăţeşte abilitatea de a face predicţii.
Algoritmi de clasificare şi arbori de decizie	Determină separarea naturală a datelor bazându-se pe o variabilă ţintă. Primele separări apar pe cele mai semnificative variabile. O ramură într-un arbore de decizie poate fi văzută ca o parte condiţională dintr-o regulă. Cei mai cunoscuţi algoritmi sunt Classification and regression trees (CART) sau CHi-squared Automatic InDuction (CHAID).
Reţele neuronale	Utilizează o colecţie de variabile de intrare, funcţii matematice şi ponderi ale intrărilor pentru a estima valoarea variabilelor ţintă. Prin intermediul unui ciclu iterativ de antrenare, o reţea neuronală îşi modifică ponderile până când rezultatul estimat se potriveşte valorilor actuale. Odată antrenată, reţeaua este un model care poate fi utilizat pentru noi date în scopuri predictive.
Algoritmi genetici	Folosesc un proces iterativ de operaţii de selecţie, încrucişare şi mutaţie pentru a elabora generaţii succesive de modele. Este utilizată o funcţie de potrivire pentru a păstra unii membrii şi a renunţa la alţii. Algoritmii genetici sunt utilizaţi în principal pentru a optimiza topologiile de reţele neuronale şi ponderile. Ei înşişi pot fi utilizaţi pentru modelare.

4. Aplicaţii ale mineritului de date şi tehnologii de minerit de date

Există, în principal, trei tipuri de probleme care pot fi abordate cu mineritul de date:

Clasificarea şi Regresia În aceste aplicaţii, mineritul de date produce modele predictive care plasează obiectele în grupuri de clasificare sau valori atribuite. Modelul de la aplicaţia de clasificare plasează un obiect într-un grup de clasificare, în timp ce un model de regresie atribuie o valoare unui obiect. De exemplu, modelul predictive „femeile având vârsta cuprinsă între 20 şi 30 de ani, care trăiesc în mediul urban răspund bine la reclamele primite prin poştă", este un model provenit de la o aplicaţie de clasificare, în timp ce modelul predictiv, „femeile având vârsta cuprinsă între 20 şi 30 de ani, care trăiesc în mediul urban răspund 70% din timp la reclamele primite prin poştă" este un model provenit de la o aplicaţie de regresie. De observat că aplicaţiile de clasificare şi regresie se pot transforma cu uşurinţă dintr-una într-alta prin înlocuirea valorii numerice cu o frază descriptivă şi invers. Un alt mod de a gândi aplicaţiile de regresie este acela de a imagina mulţimi de date cantitative, care pot fi descrise prin formule matematice; aplicaţiile de minerit de date utilizând regresia caută să descopere formule matematice care descriu datele. O dată ce formula a fost găsită, date noi pot fi introduse în formulă şi, în acest caz, predicţia este uşor de făcut. Tehnicile de calcul şi statistice care sunt utilizate în aplicaţii de clasificare şi regresie includ: arbori de decizie, reţele neuronale, naïve-Bayes şi K-nearest neighbor.

Asociere şi Secvenţiere Cunoscute şi ca analiza coşului de piaţă, în aceste aplicaţii mineritul de date produce modele prescriptive. De exemplu, un model prescriptiv cum ar fi „boiaua şi pastele sunt adesea cumpărate împreună" este un model provenit de la o aplicaţie de asociere sau secvenţiere. Tehnicile de calcul utilizate în aceste tipuri de aplicaţii sunt reprezentate de un algoritm de numărare. Secvenţierea adaugă analiza în timp într-un astfel de algoritm.

Clusterizare În aceste aplicaţii, mineritul de date produce tot modele descriptive. Aceste modele descriptive grupează obiectele şi le exclude pe cele care nu se aseamănă. De exemplu „femeile având vârsta cuprinsă între 20 şi 30 de ani, care trăiesc în mediul urban, au o anume afinitatea de a se autoapăra la produse; cele care trăiesc în mediul rural nu o au" este un posibil model descriptiv provenit de la o aplicaţie de clusterizare. Tehnicile de calcul şi statistice care sunt folosite în aplicaţii de clusterizare conţin o componentă subiectivă, care necesită implicarea unui expert uman.

5. Produse comerciale de minerit de date

Selectarea unui produs comercial de minerit de date este un proces complex. Mai întâi, alegerea unui tip de aplicaţie de minerit de date o dată cu determinarea tipurilor de informaţii care se doresc a fi descoperite prin mineritul de date merg mână în mână. De exemplu, dacă se doreşte să se determine profitabilitatea unui client, s-ar putea alege o aplicaţie de regresie. Dacă se doreşte descoperirea de corelaţii între datele aflate în câmpuri diferite din baze de date (de exemplu, tipuri de simptome care apar, de obicei, în asociere cu diferite boli) s-ar putea alege un algoritm de asociere. Odată ales tipul de aplicaţie de minerit de date, este necesar să se selecteze instrumentele de minerit de date, care utilizează algoritmii potriviţi, care vor urmări cel mai bine cerinţele mineritului. Adesea, este recomandată părerea unui expert în acest proces de decizie, în principal, datorită complexităţii algoritmilor disponibili şi a confuziei create de abundenţa de produse de minerit de date, aflate pe piaţă. Un număr considerabil de firme par a fi specializate în afaceri în analizarea şi compararea mai multor produse de minerit de date pentru clienţi.

Faptul că diferite aplicaţii de minerit de date sunt aplicabile numai la diferite tipuri de scopuri conduc în mod natural la o specializare a produselor comerciale de minerit de date pe diferite tipuri de afaceri. Vânzătorii de produse de minerit de date de cele mai multe ori îşi proiectează aplicaţiile în mod diferenţiat. Deoarece la diferite niveluri pot să existe tipuri diferite de afaceri, produsele comerciale de minerit de date au fost proiectate să răspundă cerinţelor atât ale depozitelor mici de date, dar şi ale masivelor de date.

Piaţa de produse de minerit de date este foarte complicată deoarece sunt în continuu dezvoltate produse de minerit de date noi, iar instrumentele existente sunt în continuu îmbunătăţite. Aceasta face ca utilizatorii să caute în permanenţă şi să încerce produse noi dar, în acelaşi timp să actualizeze produsele existente. De exemplu [departamentul pentru] managementul datelor de la Boston-based Fleet Bank şi grupul de analiză au testat cel puţin 15 instrumente [de minerit de date], inclusiv online analytical processing (OLAP) analitice în doi ani de existenţă, declara Victor Hoffman, vicepreşedintele şi şeful grupului de analişti de la Fleet.

În timp ce piaţa de produse comerciale de minerit de date este dezarmant de complexă, un punct bun ar putea fi în strânsă legătură cu existenţa sistemelor de management de baze de date relaţionale: majoritatea vânzătorilor de sisteme de gestiune

de baze de date au recunoscut cererea crescută de aplicații de minerit de date și și-au încorporat instrumente de minerit de date în sistemelor lor. De exemplu, Oracle Corporation include *Darwin* cu *Oracle 9i* system. *Darwin* folosește rețele neuronale, clasificare și regresie, arbori de decizie și algoritmi de clusterizare, printre altele[6]. *DB2 Intelligent Miner For Data* de la IBM[7] include suport pentru un număr de aplicații de minerit de date incluzând clasificare, asociere, secvențiere și clusterizare. *SQL Server 2000* de la Microsoft[8] conține suport pentru două clase de aplicații de minerit de date: arbori de decizie și clusterizare. Dincolo de aplicațiile de minerit de date disponibile pentru unul dintre sistemele de gestiune de baze de date relaționale există o mare varietate de instrumente de minerit de date specifice. De exemplu, sistemele de analizare a managementului relațiilor cu clienții (customer relationship management = CRM) include *Clementine* de la *Cognos Corporation*[9] și *KnowledgeSever for CRM* de la *Angoss Corporation*[10]. Produsele de minerit de date *Ultragem* folosesc algoritmi genetici pentru analizarea bazelor de date de clienți și produse. Un sector special al produselor de minerit de date este cel dedicat domeniului financiar. De exemplu, *Institutul SAS* are un produs care poate fi utilizat pentru a ajuta analiza și managementul riscurilor asociate creditului pentru instituțiile financiare[11]. Mineritul de date beneficiază de extensii serioase în sectoarele bazelor de date științifice: există multe produse de minerit de date în domeniul biomedical și al analizelor bioinformatice. De exemplu, produsul de minerit de date de la *SciTegic*[12], numit *Pipeline Pilot* a fost dezvoltat special pentru descoperirea de medicamente, ca și produsele de minerit de date de la *Anvil Informatics*[13].

6. Concluzii

Una dintre definițiile acceptate pe scară largă a termenilor **minerit de date** și **descoperire de cunoștințe** este dată de Fayyad et al. [5]:

„**Mineritul de date/descoperirea de cunoștințe** *reprezintă procesul netrivial de identificare de șabloane de date inteligibile, valide, noi, potențial utilizabile*".

Tehnologia dedicată **descoperirii de cunoștințe** contribuie serios la dezvoltarea următoarei generații de sisteme informatice și sisteme de gestiune de baze de date prin capabilitățile ei de a extrage informație nouă înglobată în baze de date eterogene de volum mare și de a construi cunoștințe.

Un proces de **descoperire de cunoștințe** include „*selectare de date din depozite mari de date, curățare, preprocesare, transformare și reducere, data mining, selecție (sau combinare) de model, evaluare și interpretare și consolidarea și utilizarea cunoștințelor extrase*" [8]. În particular, **mineritul de date** se ocupă cu dezvoltarea algoritmilor de extragere de noi șabloane din statistici, modelare cu rețele neuronale și vizualizare pentru clasificarea datelor și identificare de șabloane.

Descoperirea de cunoștințe are ca scop crearea mecanismelor prin care informația este transformată în cunoștințe prin intermediul ipotezelor de testare și formalismelor teoretice.

Din definiția termenului de **minerit de date** pot fi observate următoarele aspecte:

1. **Mineritul de date** nu este o analiză simplă și nici nu este în mod necesar egală cu machine learning. **Mineritul de date** nu este echivalent cu extragerea de date. Nu este trivial de menționat că seturile de date considerate sunt *mari*. În cele mai multe cazuri, este posibilă o analiză statistică exhaustivă și este de dorit ca ea să fie cât mai riguroasă (multe metode din mineritul de date conțin un grad de nedeterminism care le permite scalarea la seturi de date masive).
2. Unul dintre aspectele necunoscute la începutul procesului și care trebuie găsit, constă în faptul că **mineritul de date** nu se aplică în cazul problemelor deterministe sau deductive. **Mineritul de date,** în sens larg, ar putea fi o activitate *abductivă* (numită *hypothesis* de filozoful și logicianul C.S. Peirce (1878)) care nu acoperă simultan o structură oarecare în cadrul datelor și o ipoteză de explicare a ei. Aceasta va necesita structuri conceptuale sofisticate prin care o ipoteză poate fi reprezentată într-o mașină. În cadrul *descoperirii de cunoștințe*, accentul pare să se pună pe metode inductive de învățare, unde scopul este acela de a construi un model pentru intensitatea[14] unei categorii oarecare din exemplele de instruire. Deoarece structura este

---

[6] http://www.oracle.com/ip/analyze/warehouse/datamining/
[7] http://www-4.ibm.com/software/data/iminer/fordata/about.html
[8] http://www.microsoft.com/sql/productinfo/datamine.htm
[9] http://www.cognos.com
[10] http://www.angoss.com
[11] http://www.cio.com/archive/051598_mining.html
[12] http://www.scitegic.com/products_services/pipeline_pilot.htm
[13] http://www.anvilinformatics.com/
[14] intensitatea se referă în acest caz mai degrabă la descrierea generală a unei categorii decât la exemplele ei specifice

cunoscută în sens larg, aceasta nu se referă la acțiunea de *mining per se*, ci mai degrabă la o formă de *descoperire de cunoștințe*. O singură excepție apare atunci când exemplele de instruire, ele însele, sunt doar o ipoteză, generată mai degrabă din date decât *a priori*, într-un efort de stabilire a claselor cu care se reprezintă datele, așa cum este cazul unor instrumente ca AutoClass [3].

3. Structura descoperită trebuie să fie validă, adică trebuie arătat că poate fi o inferență semnificativă sau fezabilă cu un grad oarecare de confidențialitate. Metricile de fiabilitate sunt cerute ca suport al ipotezelor prezentate, dar și pentru a diferenția semnificativul din marginal sau irelevant.
4. Constatările pot fi de domeniul noului. Mașina nu are o imagine asupra a ceea ce este cunoscut sau nu de către experți, adică nu are metode de a mapa noutatea pe domeniul discursului. Prin urmare, este posibil să se postproceseze rezultatele astfel încât majoritatea inferențelor similare să fie grupate laolaltă într-o formă generalizată numită meta-instruire [1].
5. Structura descoperită trebuie să fie utilă, adică să fie explicabilă și aplicabilă într-o manieră care are sens în contextul domeniul de aplicare curent. Seturile mari de date pot să conțină foarte multă structură care ea însăși nu este utilă și orientarea efortului pe aceste părți care sunt interesante este problematică deoarece este prin definiție necunoscută la început.

De obicei, **mineritul de date** se referă la cazurile în care datele sunt prea mari sau complexe pentru a permite fie o analiză manuală, fie o analiză prin metode de interogare simple.

Mineritul de date are două etape principale:

- preprocesarea datelor, în timpul căreia caracteristicile relevante de nivel înalt sau atributele sunt extrase din date de nivel scăzut și
- recunoașterea de forme (pattern recognition), în care recunoașterea unui șablon de date se face prin intermediul acestor caracteristici (figura 3).

Preprocesarea datelor este de cele mai multe ori consumatoare de timp, prin urmare este critică. Pentru a asigura succesul în procesarea din **mineritul de date** este important ca toate caracteristicile extrase din date să fie relevante pentru problemă și reprezentative ca date.

În funcție de tipul datelor care trebuie „minerite", pasul de preprocesare constă în mai multe acțiuni. Dacă dimensiunea datelor este foarte mare, va trebui să se recurgă la simplificare, să se lucreze cu foarte puține instanțieri sau să se utilizeze tehnici multirezoluție și să se lucreze cu datele la o rezoluție macrogranulară. Apoi zgomotul este înlăturat, pe cât posibil, și sunt extrase caracteristicile relevante. În unele cazuri, în care sunt disponibile date din diferite surse sau senzori, este necesară fuziunea datelor pentru a permite exploatarea tuturor datelor disponibile pentru problemă. La sfârșitul acestui prim pas, se obține un vector caracteristic pentru fiecare instanțiere. În funcție de problemă și de date, s-ar putea să fie necesară reducerea numărului de caracteristici folosind tehnici de selecție sau de reducere a dimensiunii cum ar fi analiza componentei principale [13] sau a versiunilor sale non-lineare. După această preprocesare, datele sunt gata pentru detectarea de șabloane prin intermediul algoritmilor de clasificare, clustering, regresie etc. Aceste șabloane îi sunt afișate utilizatorului în vederea validării.

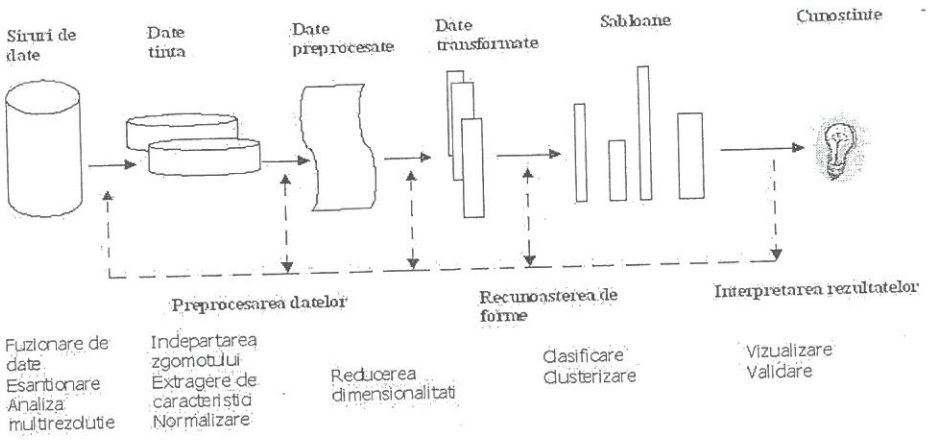

Figura 3: Mineritul de date – un proces iterativ și interactiv
(preluată de la Erick Cantu-Paz and Chandrika Kamath, "*On the use of evolutionary algorithms in data mining*")

Mineritul de date este un proces iterativ și interactiv. Ieșirea din oricare dintre pași sau feedback-ul de la experți poate să conducă la o rafinare iterativă a unora sau a tuturor acestor activități.

Aşa cum menţionează Erick Cantu-Paz şi Chandrika Kamath, în "*On the use of evolutionary algorithms in data mining*" există unele dezbateri legate de definirea termenului de **minerit de date** în care majoritatea practicienilor şi a celor care au propus definiţii au căzut de acord că **mineritul de date** este un domeniu multidisciplinar, împrumutând idei de la machine learning şi inteligenţa artificială, statistică, procesare de semnal şi imagine, calcul de înaltă performanţă, optimizare, recunoaştere de forme etc. Ca element de noutate, trebuie remarcată confluenţa ramurilor mature ale acestor tehnologii, la un moment dat, în care ele trebuie exploatate în analiza masivelor de date.

Majoritatea produselor de minerit de date sunt orientate mai mult pe tehnologie, decât pe uşurinţa în utilizare, scalabilitate sau portabilitate. În acelaşi timp, există numeroase estimări făcute de organizaţiile de standardizare şi consorţii care au căzut de acord asupra unui mod standardizat de a utiliza mineritul de date împreună cu produsele actuale de management de date cum sunt bazele de date SQL şi depozitele de date. Trebuie menţionate aici trei aspecte importante:

1. ISO/IEC JTC1 SC32 WG4: SQL/MM Part 6 Data Mining A collection of SQL user-defined types and routines to compute and apply data mining models.
2. The Data Mining Group (DMG): Predictive Model Markup Language (PMML) An XML based specification for data mining models.
3. OMG: Common Warehouse Metamodel (CWM): Chapter 14 Data Mining A UML/XML based specification for data mining metadata.

Toate cele trei aspecte provin din tehnologiile de minerit de date, specifice diferitelor domenii, şi se ocupă de problema uşurinţei în utilizare prin ascunderea complexităţii algoritmilor fundamentali de minerit de date[15] (Friedemann Schwenkreis). Extensiile de minerit de date definite în SQL/MM merg chiar mai departe introducând rutinele SQL care permit invocarea funcţiilor de minerit de date ca parte a instrucţiunilor SQL. Implementarea SQL (optimizatorul) ţine sub control execuţia şi decide dacă este utilizat sau nu paralelismul şi unde au loc calculele.

Standardele introduc cerinţe de nivel înalt asupra produselor de minerit de date, cerinţe care provin, în principal, de la utilizatorii de tehnologie de minerit de date. Se pare că, în cazul mineritului de date, standardele nu au numai intenţia de a unifica produsele existente cu o funcţionalitate bine cunoscută, ci şi de a proiecta (parţial) funcţionalitatea astfel încât viitoarele produse să se potrivească mai bine cerinţelor din lumea reală. Acest aspect poate fi privit ca o tendinţă generală în eforturile actuale de standardizare. Obiectivul este, mai degrabă, acela de a avea o specificaţie standardizată cât mai devreme posibil, decât de a defini un standard după ce majoritatea produselor au adoptat deja un standard "de facto". Prin urmare, din acest punct de vedere, noile abordări ale produselor sunt conduse de standarde, şi nu standardele de produse.

„ ...Pentru a concluziona, instrumentele de minerit de date sunt orientate către a obţine abilităţi de analiză multimedia şi capabilităţi de analizare simultană a numeroase tipuri de baze de date. Utilizatorii finali sau specialiştii în mineritul de date vor fi persoane care operează cu instrumentele care sunt înglobate în pachetele software standard, cel mai probabil utilizând tehnologii de reţele neuronale"[16] [16].

Deci, mineritul de date este un domeniu în plină dezvoltare al ştiinţei calculatoarelor, care va furniza un nivel nou şi eficient de informaţii şi de descoperire de cunoştinţe de care vor beneficia toţi utilizatorii din domeniul memorării computerizate de date.

---

[15] http://research.microsoft.com/~jamesrh/hpts2001/submissions/FriedemannSchwenkreis.htm

[16] "*... To conclude, data mining tools are heading in the direction that they will have multi-media analysis abilities, and be able to analyze several types of databases simultaneously. The end-users or professional data miners will be the people operating the tools that are embedded in to standard software packages, most probably utilizing neural network technology.*"

Bibliografie

1. **BRADSIL, P. B., K. KONOLIGE, K.** (Eds.): Meta-Learning, Meta-Reasoning and Logics, Boston, MA, USA, Kluwer Academic Press, 1990.
2. **CANTU-PAZ, E., CHANDRIKA KAMATH**: On the Use of Evolutionary Algorithms in Data Mining; http://www.llnl.gov/CASC/sapphire/pubs/heuristic.pdf
3. **CHEESEMAN, P., J. STUTZ**: Bayesian Classification: Theory and results. În: Eds. Fayyad, U., Piatetsky-Shapiro, G, Smyth, P. and Uthurusamy, R. (Eds.) Advances in Knowledge Discovery and Data Mining, Cambridge, MA: AAAI/MIT Press, 1996.
4. **COLLIER, K. B., E. CAREY, C. GRUSY, D. MARJANIEMI, D. SAUTTER**: A Perspective on data Mining, Northern Arizona University, July 1998, insight.nau.edu/downloads/DM%20Perspective%20v2.pdf
5. **FAYYAD, U., G. PIATETSKY-SHAPIRO, P. SMYTH**: From Data Mining to Knowledge Discovery in Databases. În: AI Magazine, Fall 1996a.
6. **FAYYAD, USAMA, G. PIATETSKY-SHAPIRO, P. SMYTH**: From Data Mining to Knowledge Discovery: An Overview: În: Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996b.
7. **FAYYAD, USAMA, G. PIATETSKY-SHAPIRO, P. SMYTH:** Knowledge Discovery and Data Mining: Towards a Unifying Framework. În: KDD-96 Conference Proceedings, ed. E. Simoudis, J. Han, and U. Fayyad, AAAI Press, 1996c.
8. **FAYYAD, U.**: Editorial. Data Mining and Knowledge Discovery, 1997.
9. **FRAWLEY, W., G. PIATETSKY-SHAPIRO, C. MATHEUS**: Knowledge Discovery in Databases: An Overview. În: AI Magazine , Fall 1992, pp 213-228.
10. **FRIEDEMANN SCHWENKREIS** editor of ISO/IEC 13249-6 SQL/MM Data Mining IBM Deutschland Entwicklung GmbH http://research.microsoft.com/~jamesrh/hpts2001/submissions/FriedemannSchwenkreis.htm
11. **HAND, D., H. MANNILA, P. SMYTH**: Principles of Data Mining. MIT Press, Cambridge, MA, 2001
12. *** * ***: IBM's Data Mining Technology – White Paper data Management Solutions, April, 1996 (http://cgi.di.uoa.gr/~rouvas/research/ibm/IBM-datamine.html)
13. **JACKSON, J.E.**: An User's Guide to Principial Components, New York, NY:John Wiley, 1991.
14. **MENZIES, T., Y. HU**: Data Mining For Very Busy People. În: IEEE Computer, October 2003, pp 18-25.
15. *** * ***: NASA Workshop on Issues in the Application of Data Mining to Scientific Data, October 19 – 21, 1999, University of Alabama in Huntsville, Huntsville, Alabama, www.cs.uah.edu/~thinke/NASA_Mining/DMFinalReport.pdf
16. **PATEL, L., S. PERRY, D. TAYLOR, A. BROWN, S. TIKE**: The Future of Data Mining http://www.aston.ac.uk/~golderpa/CS342/grouppages/dssg9/fut.html
17. **PEIRCE, C. S.**: Deduction, induction and hypothesis. În: Popular Science Monthly, 13, 1878.
18. **SHAPIRO, G.** Editor, http://www.megaputer.com/dm/index.php3
19. **TAIPALE, K. A.**: Data Mining and Domestic Security: Connecting the Dots to Make Sense of Data (http://www.stlr.org/cite.cgi?volume=5&article=2), Center for Advanced Studies in Science and Technology Policy (http://www.advancedstudies.org/). 5 Colum. Sci. & Tech. L. Rev. 2 (December 2003).
20. *** * ***: Wikipedia article „*Data mining*", http://en.wikipedia.org/wiki/Data Mining
21. **MAFRUZ ZAMAN ASHRAFI, D. TANIAR, K. A. SMITH**: A Data Mining Architecture for Clustered Environments, Publisher: Springer-Verlag Heidelberg, ISSN: 0302-9743
22. http://www3.shore.net/~kht/text/dmwhite/dmwhite.htm
23. http://www.cio.com/archive/051598_mining.html
24. http://www.oracle.com/ip/analyze/warehouse/datamining/
25. http://www-4.ibm.com/software/data/iminer/fordata/about.html
26. http://www.cognos.com/
27. http://www.angoss.com/
28. http://www.microsoft.com/sql/productinfo/datamine.htm
29. http://www.scitegic.com/products_services/pipeline_pilot.htm
30. http://www.anvilinformatics.com/
31. http://en.wikipedia.org/wiki/Data Mining