DATA MINING – TEHNOLOGII DEDICATE EXTRAGERII CUNOSTINTELOR

Obiective:
- însusirea tehnologiei Data Mining de extragere a cunostintelor din colectiile de date existente;
- însusirea unor tehnici Data Mining pentru obtinerea unor solutii în cadrul problemelor decizionale.

Concepte cheie: Data Mining; tehnici Data Mining.

Existenta unor volume imense de date a pus problema reorientării utilizării lor de la un proces de exploatare retrospectiv către unul prospectiv. Data Mining poate avea mai multe definitii, însă toate converg în esentă către miezul problemei, si anume că acest concept reprezintă un proces de extragere de informatii noi din colectiile de date existente. Termenul de dată are semnificatia de descriere a unui eveniment bine determinat care se produce în lumea reală si este perfect verificabil.

Prin tehnologia Data Mining se prelucrează date care referă perioade anterioare (date istorice), care sunt examinate si sunt deja cunoscute, pe baza lor constituindu-se un model. Acest model va putea fi aplicat situatiilor noi de acelasi tip cu cele deja cunoscute. Informatiile care se pot obtine prin Data Mining sunt predictive sau descriptive. De exemplu directionarea actiunilor de marketing pot constitui o problemă tipică predictivă. Detectarea fraudelor produse cu carduri bancare reprezintă o problemă tipică de aplicatie descriptivă.

Dezvoltarea tehnicilor de Data Mining se explică prin acumularea de volume pe care acestea le-au derulat de-a lungul anilor. De asemenea, concurenta tot mai acerbă precum si cresterea exigentelor pietei au determinat firmele să ia tot mai mult în considerare potentialul urias pe care îl oferă arhivele de date. Alături de arhivele de date memorate pe suporturi informatice mai există încă doi factori care au dus la necesitatea Data Mining: existenta si perfectionarea algoritmilor si a produselor program dedicate precum si cresterea capacitătii de memorare si prelucrare a calculatoarelor care permit tratarea corelativă a volumelor mari de date.

Este de remarcat că depozitele de date pot fi surse pentru Data Mining, iar rezultatele obtinute pot completa câmpurile înregistrărilor din depozitele de date, care apoi pot fi valorificate prin proiectiile multidimensionale specifice OLAP.

Potentialul oferit de Data Mining se încorporează în procesele comerciale ale firmelor, iar căutarea informatiilor nu devine un scop în sine ci este utilă doar dacă este transformată ca actiune. Astfel firmele pot alege să reactioneze sau nu la situatiile diverse create de realitate (diminuarea numărului de clienti, scăderea vânzărilor, pierderea unor piete de desfacere etc.). Pasul următor după această alegere este exploatarea propriu-zisă a datelor utilizând diversi algoritmi.

De multe ori, actiunea de Data Mining poate fi un esec si nu o reusită, fiind posibil ca măsurile luate să nu fie adecvate informatiilor obtinute.

Toate elementele considerate anterior conduc spre ideea de ciclu în utilizarea Data Mining în cursul căruia sunt patru etape:
- definirea oportunitătilor comerciale si a datelor
- obtinerea de informatii din colectiile de date existente prin tehnici Data Mining;
- adoptarea deciziilor si actiunilor în urma informatiilor rezultate;

- cuantificarea cât mai corectă a rezultatelor concrete pentru a identifica si alte căi de exploatare a datelor.

Căutarea cunostintelor si verificarea ipotezelor

Tehnicile de Data Mining se pot aplica atât ascendent, cât si descendent. Pentru abordarea descendentă se iau în considerare ipotezele formulate în prealabil prin alte mijloace. Abordarea ascendentă urmăreste extragerea de cunostinte sau informatii noi din date disponibile, această căutare putând fi dirijată sau nedirijată.

Căutarea dirijată presupune că se ia în considerare un atribut sau un câmp, ale cărui valori se explică prin celelalte câmpuri. Căutarea nedirijată identifică relatiile sau structurile din datele examinate fără a asigura prioritate unui câmp sau a altuia. Ceea ce se exploatează prin Data Mining sunt colectii de date constituite pentru alte scopuri (exemplu tranzactii derulate pe o perioadă de timp). Deseori la acest tip de date se adaugă si cele provenite din alte surse cum statistici oficiale care privesc evolutia în ansamblu a economiei, date privind concurenta sau măsuri legislative. De aceea se foloseste tot mai des notiunea de informatie ascunsă în sensul că este aproape imposibilă detectarea corelatiilor sau raporturile pe care datele le încorporează în mod intrinsec.

Rezultatele obtinute sunt cu atât mai relevante cu cât ele se bazează pe un volum mare de date. Datele pot fi exploatate pentru a obtine informatii prin diverse tehnici cum sunt: retele neuronale, arbori de decizie, algoritmi genetici, analiza grupurilor, rationamente bazate pe cazuri, analiza legăturilor. Aceste tehnici pot fi asociate cu tehnici statistice cum sunt regresiile sau analiza factorială. Data Mining nu este capabilă, ca tehnică, să rezolve orice problemă de
gestiune. De fapt ceea ce poate oferi se rezumă la câteva actiuni cum sunt: clasificarea, estimarea, predictia, gruparea, analiza grupărilor, care folosite la locul potrivit pot deveni utile pentru o multime de probleme din domeniul decizional.

Destinatia si caracteristicile actiunilor oferite de Data Mining

Clasificarea are ca scop plasarea obiectelor prelucrate într-un grup limitat de clase predefinite. De exemplu, vânzarea unui produs nou se poate încadra într-una din următoarele categorii de risc: scăzut, mediu, ridicat. Obtinute în mod clasificat vor fi reprezentate sub formă de înregistrări care la rândul lor sunt
compuse din atribute sau câmpuri. Ca tehnici de Data Mining pentru clasificare sunt arborii de decizie si rationamentul bazat pe cazuri.

Estimarea va atribui o valoare unei variabile pe baza celorlalte date de intrare. Rezultatele obtinute în urma estimării sunt valori continue. Pentru acest tip de prelucrări se pot utiliza retelele neuronale.

Predictia poate clasa înregistrările luate în considerare în functie de un anumit comportament sau o valoare viitoare estimată. De aceea se va recurge la o colectie de exemple care vizează date din trecut, în care valorile variabilei de previzionat sunt deja cunoscute. Cu ajutorul lor se va construi un model care va putea explica comportamentul observat. Aplicând acest model înregistrărilor care fac obiectul prelucrării, se va obtine o predictie a comportamentului sau a valorilor acestora în viitor.

Gruparea poate duce la determinarea acelor obiecte care apar cel mai frecvent împreună. Un exemplu este „analiza cosului gospodăriei” în evaluările

statistice.

Analiza grupului urmăreste o dividere a populatiei eterogene în grupuri mai omogene, care poartă numele de clustere.

În această tehnică nu se pleacă de la un set predeterminat de clase si nici din exemple din trecut. Segmentarea pe grupuri se face în functie de similitudinile obiectelor.

Explorarea datelor – continut si etape

Programele care realizează implementarea algoritmilor pentru Data Mining nu sunt suficiente. Ele trebuie alimentate cu date care provin din diverse surse organizate pentru alte scopuri. De aceea este necesar un proces de curatare a acestora si de uniformizare pentru a fi explorate asa cum sunt ele furnizate de programe, continutul lor trebuind a fi analizat de specialisti care vor identifica informatiile utile pe care acestea (rezultatele) le conŃin. Având în vedere aceste particularităti, tehnicile de Data Mining se pot utiliza numai în procese specifice complexe si de cele mai multe ori neliniare. Se pot astfel distinge etapele:

- definirea problemei;
- identificarea surselor de date;
- colectarea si selectarea datelor;
- pregătirea datelor;
- definirea si construirea modelului;
- evaluarea modelului;
- integrarea modelului.

Definirea problemei constă în sesizarea unei oportunităti sau necesităti de afaceri. De aceea se va delimita ceea ce urmează a fi rezolvat prin Data Mining, obiective urmărire si rezultate scontate. Problema ce urmează a fi rezolvată prin Data Mining este o parte componentă a oportunitătii organizatiei, dar nu se identifică cu ea. De asemenea problema trebuie să primească o formă adecvată pentru a putea fi tratată cu această tehnică.

Identificarea surselor de date constă în stabilirea structurii generale a datelor necesare pentru rezolvarea problemei, precum si regulile de constituire a acestora si localizarea lor. Fiecare sursă de date va fi examinată pentru o familiarizare cu continutul său si pentru identificarea incoerentelor sau a problemelor de definire.

Colectarea si selectia datelor este etapa în care se face extragerea si depunerea într-o bază comună a datelor care urmează a fi utilizate ulterior.
Această etapă ocupă un timp mare, cam 80% din timpul total, iar existenta depozitelor de date constituie un real avantaj.

În functie de limitele echipamentelor de calcul folosite, de produsele program aplicate colectiilor de date si nu în ultimul rând de bugetul disponibil se poate prelucra întregul fond de date disponibil sau un esantion. Dacă optiunea aleasă este dirijată spre lucrul cu esantionare, atunci trebuie respectate toate regulile si cerintele de selectare a acestora.

Pregătirea datelor. Datele sunt de obicei stocate în colectii de date care au fost construite pentru alte scopuri. De aceea firesc este să existe o fază preliminară de pregătire înainte de extragere prin Data Mining. Transformările la care sunt supuse datele pentru Data Mining se referă la: valori extreme, valori lipsă, valori de tip text, tabele.

Tratarea valorilor extreme se poate face prin încadrarea între anumite limite cuprinse între medie si un număr de abatere standard prin excludere sau limitare sau prin izolarea vârfurilor.

În cazul valorilor lipsă se pot elimina câmpurile cu valori nule din înregistrări, sau se pot completa câmpurile cu date de valori medii, deoarece existenta lor poate duce la o functionare incorectă a algoritmilor de Data Mining.

Valorile de tip text ridică probleme întrucât separarea prin spatii a cuvintelor duc la aparitia de valori diferite. Din acest motiv este indicată eliminarea lor, dar dacă prelucrarea lor nu poate fi eliminată, solutia cea mai pertinentă este de codificare prin tabele de corespondente, în care să se evidentieze toate sirurile valide de caractere.

Rezumarea se aplică atunci când datele sunt considerate a reprezenta detalii nesemnificative pentru rezolvarea problemei, sau când numărul de exemple este insuficient.

Codificarea incoerentă apare în momentul în care obiecte identice sunt reprezentate diferit în unele din sursele utilizate. Incompatibilitătile arhitecturale informatice se referă la diferentele existente între modul de reprezentare internă a valorilor datorat creării lor cu sisteme din generatii diverse.

Definirea si construirea modelului este etapa care se apropie cel mai mult de notiunea de Data Mining si se referă la crearea modelului informatic care va efectua exploatarea. Etapa de definire si construire a modelului este însotită de faza de instruire sau învătare, depinzând de tehnicile de Data Mining utilizate.

Indiferent de aceste tehnici toate au de parcurs două etape: învătarea si testarea. Învătarea presupune existenta unui set suficient de reprezentativ de exemple complete de la care se porneste pentru a identifica relatiile de legătură între valorile câmpurilor sau atributelor. Se consideră ca fiind încheiat procesul de învătare, în momentul în care rezultatele obtinute prin model se apropie suficient de mult de solutiile continute de datele după care s-a învătat. Nu întotdeauna rezultatele sunt cele scontate si atunci modelul va fi supus testării cu date diferite de cele folosite pentru învătare, dar care aparŃin aceleiasi colecŃii. În această etapă sunt formulate alte două obiective, si anume: obtinerea de date preclasate si distribuirea acestora în seturi de învătare, testare sau evaluare.

Evaluarea modelului are ca scop de a determina corect valorile în care modelul are capacitatea de a determina corect valorile pentru cazurile noi. Modelul va fi astfel aplicat asupra ultimei părti din datele preclasate care sunt dedicate evaluării. Procentul de eroare ce se stabileste acum va fi considerat că va fi acceptat si pentru datele noi.

Performantele unui model se vor aprecia cu „matricea de confuzie” care are rolul de a compara situatia reală cu cea pe care modelul o furnizează. Integrarea modelului este etapa în care se finalizează procesul, prin încorporarea modelului în SIAD ca element de bază, sau prin includerea sa întrun proces decizional general din organizatie.

Rationamentul bazat pe cazuri

Prin această tehnică se caută o rezolvare a problemelor apărute prin analogie cu experienta acumulată. Această metodă se poate aplica pentru clasificări si pentru predictii. Cazurile pe care este bazat rationamentul sunt memorate ca înregistrări

compuse din setul de atribute care descriu fiecare caz. Un caz nou este prezentat tot ca o înregistrare, numai că în câmpurile în care valoarea trebuie determinată sunt vide. Pentru a determina aceste valori se caută înregistrările cu care înregistrarea „caz nou” se aseamănă si continutul acestora se consideră a fi răspunsul. Prin urmare se poate afirma că există două functii fundamentale de prelucrare:
a) măsurarea distantei dintre membrii fiecărui cuplu de înregistrări, pentru a afla vecinele cele mai apropiate;
b) combinarea rezultatelor obtinute de la „vecine” în răspunsul propus pentru cazul curent.

Măsurarea distantei dintre câmpuri. Se numeste distantă expresia modului în care se evaluează similitudinea. Distanta are ca proprietăti: poate fi definită si se prezintă ca un număr real; distanta de la un element la el însusi este totdeauna nulă; sensul de măsurare este fără semnificatie în maniera că distanta de la elementul A la elementul B este egală cu distanta de la B la A si nu există un punct C intermediar lui A si B prin a cărei parcurgere să se scurteze drumul de la A la B.
Ca moduri de calcul pentru distanta câmpurilor numerice se enumeră:
- diferenta între valoare absolută |A-B|;
- pătratul diferentei (A-B)2;
- diferenta între valoare absolută normalizată |A-B| (diferenta maximă). Ultima variantă produce rezultate cu valori cuprinse între 0 si 1. Măsurarea distantei între înregistrări. Când apare necesitatea de a considera simultan mai multe câmpuri ale înregistrării, se calculează distanta pentru fiecare câmp în parte, iar rezultatul se combină într-o valoare mică care reprezintă distanta înregistrării respective.
Se vor enumera câteva procedee de combinare a distantei câmpurilor: însumarea, însumarea normalizată (suma distantelor/suma maximă), distanta euclidiană (rădăcina pătrată din suma pătratelor distantelor). Distanta euclidiană evidentiază cel mai bine înregistrările pentru care toate câmpurile sunt vecine. Combinarea rezultatelor presupune aflarea celor mai apropiati vecini, iar solutia problemei se obtine prin combinarea răspunsurilor obtinute de la acestia.

Fiecare vecin poate avea diverse variante de răspuns, dar se vor lua în calcul doar cei care sunt mai apropiati. Rezultatul ce obtine majoritatea va fi atribuit cazului curent. Cerinta minimă este ca numărul votantilor să fie impar, pentru a evita situatiile de nedeterminare.

Metodele care se bazează pe vot dau rezultate satisfăcătoare în situatiile în care răspunsurile asteptate sunt de tip enumerativ. O altă solutie posibilă este interpolarea valorilor înregistrărilor vecine care însă introduce o aplatizare a rezultatelor care se înscriu între cele două limite folosite în calcul. De asemenea, se poate constata că rezultate bune se obtin prin metode de regresie statistică aplicate asupra valorilor date de vecinii cei mai apropiati. Se obtine ecuatia unei drepte sau a unei curbe care permite calcularea mai precisă a valorilor aferente cazului curent.

Se poate concluziona că rationamentul bazat pe cazuri este o tehnică de Data Mining suficient de bună si care se poate aplica unui mare număr de probleme, caz în care conduce la solutii acceptabile. Toate acestea sunt valabile dacă volumul de date pe care se bazează este bine ales si concludent. Ca avantaje pentru această metodă se pot enumera:

- aplicarea unui mare număr de tipuri de date, pe structuri de date complexe, iar câmpurile tip text sunt mai bine tratate decât în alte tehnici;
- luarea în considerare a oricât de multor câmpuri;
- rezultatele obtinute sunt explicite;
- elementele de noutate care apar în procesul de învătare sunt usor de înglobat si de folosit în rationamente.

Ca orice metodă prezintă si unele dezavantaje dintre care se pot mentiona: volumul mare de memorie si resursă timp de prelucrare relativ mare, si de asemenea, timpul de prelucrarea mare pentru aplicarea functiilor de distantă asupra tuturor înregistrărilor si câmpurilor necesare pentru obtinerea rezultatelor.