EXPLORAREA DATELOR ŞI DESCOPERIREA CUNOŞTINȚELOR - PROBLEME, OBIECTIVE ŞI STRATEGII

Cornel Lepădatu

cornel_lepadatu@biblacad.ro

Academia Română Bucureşti

Biblioteca Academiei Române

Rezumat: Explorarea datelor şi descoperirea cunoştinţelor, „data mining", este un ansamblu de metode şi algoritmi destinat explorării şi analizei unor, adesea, mari volume de date în vederea deducerii, din aceste date, a unor reguli, a unor asocieri, a unor tendinţe necunoscute, a unor structuri specifice care să restituie în mod concis esența informației utile pentru asistarea deciziilor.

În ciuda dezvoltării rapide, domeniul data mining este încă vag definit şi lipsit de o abordare integrată, situaţie care provoacă dificultăți în procesele de predare, de învățare, de cercetare precum şi în cele de aplicare [9]. Succesul unui proiect, din orice domeniu de activitate al organizațiilor contemporane, este de multe ori compromis de propensiunea generală de a elabora soluțiile înainte de a identifica şi formula problemele.

Articolul se concentrează asupra unor aspecte importante privind natura şi calitatea datelor utilizate în aplicarea data mining, privind metodele cele mai frecvent utilizate, alegerea principalelor obiective, formularea şi tratarea problemelor în contextul strategiilor uzuale de data mining.

Cuvinte cheie: obiective data mining, problematică data mining, proces data mining, strategie data mining, tehnologie data mining.

Abstract: Data mining and knowledge discovery denote a set of methods and algorithms for exploration and analysis of (often) large volumes of data aiming to infer rules, associations, unknown trends, specific structures so that useful information may be returned in a concise form for supporting decisions. Despite its fast-paced development the data mining is still vaguely defined and lacks an integrated approach. This situation causes difficulties in teaching, learning, research and application. The success of a project in any field of activity of contemporary organizations is often compromised by the general propensity to develop solutions before identifying problems and formulating statements. The article focuses on several important aspects such as the nature and quality of data used in the application of data mining, the most commonly used methods, the choice of the main objectives, problem formulation that should be adequately addressed in the context of data mining common strategies.

Key words: data mining goals, data mining problems, data mining process, data mining strategies, data mining technology.

1. Introducere

Organizațiile au acumulat volume foarte mari de date, stocate pe suporturi informatice, privitoare la tranzacții de diverse tipuri, derulate de-a lungul multor ani. Astfel:

- băncile posedă arhive de milioane de înregistrări în care sunt consemnate în detaliu operațiile efectuate de clienții lor;
- în aproape orice firmă se găsesc mii şi sute de mii de înregistrări privitoare la cumpărările, vânzările, încasările şi plățile efectuate;
- societățile de telefonie mobilă posedă date privitoare la fiecare convorbire efectuată de abonații lor, incluzând data, momentul şi locul apelului, durata convorbirii, numărul de telefon al corespondentului;
- magazinele posedă sute de mii de înregistrări, provenind de la casele de marcaj, în care figurează nu numai articolele cumpărate ci şi cumpărătorii, identificați prin legitimațiile de acces.

O dată cu expansiunea internetului, volumul datelor stocate în format digital nu încetează să crească, din ce în ce mai rapid, peste tot în lume:

- indivizii pun, din ce în ce mai mult, informațiile pe care le dețin la dispoziția tuturor, via web;
- numeroase organizații, în special cea mai mare parte a marilor magazine, culeg din ce în ce mai multe informații despre clienții lor şi comportamentele acestora;
- foarte multe dintre procesele industriale sunt controlate informatic;
- rezultatele analizelor medicale sunt, din ce în ce mai sistematic, stocate pentru a fi analizate;
- tot mai numeroase măsurători efectuate pretutindeni în lume, ca de exemplu cele meteorologice, umplu de asemenea importante baze de date digitale.

Mijloacele şi tehnicile informatice, tot mai evoluate, au contribuit de-a lungul timpului la amplificarea capacității de memorare şi stocare a datelor iar în ultimile decenii au susținut o reorientare semnificativă, privind utilizarea volumelor de date stocate, de la un proces de explorare retrospectivă către unul cu caracter prospectiv:

- multă vreme aceste date s-au acumulat pur şi simplu în virtutea nevoii de arhivare;
- datele acumulate conțin informații şi cunoştințe „ascunse", care pot servi la bunul mers al unei organizații, dar luate ca atare, nu au mare utilitate dacă nu sunt însoțite de mecanisme care să permită explorarea lor şi înțelegerea fenomenelor care au guvernat funcționarea surselor de date;
- creşterea permanentă a concurenței, exigențele din ce în ce mai mari ale pieței au determinat organizațiile să devină conştiente de potențialul pe care aceste arhive de date îl reprezintă.

„Informația nu lipseşte, ceea ce lipseşte este timpul managerului de a considera toate informațiile care sunt disponibile" semnala încă din 1992, H. Simon, laureat al Premiului Nobel pentru economie [4]. În zilele noastre, nu numai că volumul de date stocate digital este foarte important, dar şi tipul acestor informații este foarte diversificat:

- web-ul este un exemplu, foarte prezent astăzi, de spațiu care regrupează date foarte numeroase, diverse şi variate: texte structurate sau nu, imagini, sunete, filme, etc.;
- bazele de date clienți, datele extrase din procesele de producție, rezultate ale analizelor medicale sau baze de date de măsurători mondiale pot conține de asemenea un număr important de informații eterogene: date numerice, categoriale, curbe, etc.

Există în prezent un foarte mare interes de a dezvolta tehnici care să permită utilizarea optimă a tuturor acestor stocuri de informații, pentru a extrage din ele un maximum de cunoaştere utilă:

- pe web, este vorba de a înțelege mai bine conținutul paginilor web şi cererile utilizatorilor pentru a le furniza informația țintă cea mai pertinentă posibilă şi în maniera cea mai comprehensivă posibilă;
- în cazul bazelor de date de clienți, poate fi vorba de a înțelege cât mai bine comportamentele clienților pentru a le facilita accesul la produsele care îi interesează;
- în ce priveşte datele provenite din procesele de producție, există un mare interes de a extrage din ele un maximum de cunoştințe pentru a deduce din ele bune practici de optimizare a producției;
- studiul rezultatelor analizelor medicale poate să ajute la mai buna depistare a pacienților cu risc pentru anumite boli, permițând astfel mai degrabă prevenirea decât vindecarea;
- analiza datelor meteorologice poate ajuta la mai buna înțelegere a fenomenelor generale care influențează climatul pentru a anticipa fenomenele extreme şi pentru a acționa în consecință pentru populațiile vizate.

2. Tehnologia data mining

Preocupările privind descoperirea de noi cunoştinţe utile prin analizarea de date existente au condus la dezvoltarea tehnologiei data mining ale cărei rădăcini se regăsesc în statistica matematică, în pachetele software folosite în ştiinţele sociale şi în inteligenţa artificială.

Data mining nu este nici noutate tehnologică nici ştiinţifică, metodele şi tehnicile utilizate sunt relativ vechi. Noutatea a constat în integrarea acestora în procesarea industrială a informaţiei. Dezvoltarea în decursul timpului a diverselor concepte, metode şi tehnici utilizate în prezent de tehnologia data mining se poate încadra [11] în trei perioade după cum urmează:

- *statistică* (sau preistorie): 1758, *clasificare*, Carl von Linné; 1875, *regresie liniară*, Francis Galton; 1896, *formula coeficientului de corelaţie*, Karl Pearson; 1900, *distribuţia* χ^2, Karl Pearson; 1930, *analiza factorială*, Hotteling; 1936, *analiza discriminantă*, Fisher şi Mahalanobis; 1941, *analiza factorială a corespondenţelor*, Guttman; 1943, *reţele neuronale*, Mc Culloch şi Pitts; 1944, *regresia logistică*, Joseph Berkson; 1958, *perceptronul*, Rosenblatt; 1962, *analiza datelor*, J.-P. Benzécri; 1964, *arbore de decizie AID*, J.P.Sonquist şi J.-A.Morgan; 1965, *metoda centrelor mobile*, E. W. Forgy; 1967, *metoda celor k-medii*, Mac Queen; 1972, *modelul liniar generalizat*, Nelder şi Wedderburn;
- *analiza datelor* (sau istorie): 1975, *algoritmi genetici*, Holland; 1975, *metoda de clasare DISQUAL*, Gilbert Saporta; 1977, *analiza exploratorie a datelor*, Tukey; 1980, *reţele bayesiene*, Pearl; 1980, *arbore de decizie CHAID*, KASS; 1983, *regresie PLS* (Partial Least Squares), Herman şi Svante Wold; 1984, *arbore CART*, Breiman, Friedman, Olshen, Stone; 1986, *perceptron multistrat*, Rumelhart şi McClelland; 1989, *reţele* (*auto-adaptative*), T. Kohonen;
- *explorarea datelor şi descoperirea cunoştinţelor*: ***1990*** (aproximativ), apariţia conceptului de ***data mining*****;** 1993, *arbore C4.5*, J. Ross Quinlan; 1996, *bagging* (Breiman) şi *boosting* (Freund-Shapire); 1998, *maşini cu suport vectorial*, Vladimir Vapnik; 2000, *regresie logistică PLS*, Michel Tenenhaus; 2001, *păduri aleatoare*, L. Breiman.

Principalele aspecte ale utilizării acestor tehnici, caracteristice pentru fiecare perioadă, sunt următoarele [11]:

- *statistică*: câteva sute de indivizi; câteva variabile cu datele obţinute cu o procedură specială (eşantionare, planificare experiment); ipoteze tari privind legile statistice urmate; modelele provin din teorie şi sunt confruntate cu datele; metode probabiliste şi statistice; utilizare în laborator;
- *analiza datelor*: câteva zeci de mii de indivizi; câteva zeci de variabile; construirea de tabele "indivizi × variabile"; importanţă pentru calcul şi reprezentare vizuală;
- *data mining*: mai multe milioane de indivizi; mai multe sute de variabile; numeroase variabile nenumerice, uneori textuale; date obţinute anterior studiului şi adesea în alte scopuri; date imperfecte, erori de obţinere sau de codificare, valori lipsă sau aberante; populaţie constant evolutivă (dificil de eşantionat); necesitatea unor calcule rapide, uneori în timp real; nu se caută întotdeauna optimul matematic ci modelul cel mai uşor de înţeles (sau de asimilat) de către utilizatori; ipoteze slabe privind legile statistice urmate; modelele sunt obţinute din date din care se deduc şi elemente teoretice; metode statistice, de inteligenţă artificială şi de teoria învăţării; utilizare în organizaţii.

3. Probleme data mining

Tehnologia data mining permite descoperirea de pattern-uri structurale din date utilizând algoritmi suficient de robuşti atât pentru a prelucra date imperfecte, corelate stohastic, cât şi pentru a extrage corelaţii, uneori imprecise, şi reguli utilizabile ulterior în predicţia, explicarea şi înţelegerea evoluţiei structurii datelor analizate.

În fapt, aportul data mining se rezumă la un număr limitat de acţiuni care, folosite în mod adecvat, se pot dovedi extrem de utile pentru numeroase probleme şi situaţii din domeniul decizional. Între principalele tipuri de probleme, rezolvabile cu data mining, cele mai frecvente [9, 12] sunt: *analiza asocierilor*, *pattern-uri secvenţiale*, *analiza grupurilor*, *clasificare*, *mulţimi rough*, *link mining*.

Datele disponibile sunt privite [1, 2, 3, 7, 10, 11, 13] ca reprezentând o serie de *observaţii* privind un set de caracteristici sau *variabile* $Y = \{ Y^j \mid j = 1 \div p \}$, care au fost măsurate pe un eşantion de obiecte sau *indivizi*, $X = \{ x_i \mid i = 1 \div n \}$. Există două tipuri de variabile, explicative şi de explicat: mulţimea de variabile *explicative* sau *predictive*, este constituită din variabile, fie toate cantitative, fie toate calitative, fie mixte; variabilele *de explicat* sau *de predicţie* sau *ţintă*, de asemenea, pot fi: cantitative şi calitative cu două sau mai multe modalităţi.

Analiza asocierilor. Fie $A = \{a_1, \ldots, a_j, \ldots, a_p\}$ o mulţime de articole şi fie $\boldsymbol{T} = \{\boldsymbol{t_1}, \ldots, \boldsymbol{t_i}, \ldots, \boldsymbol{t_n}\}$ o mulţime de tranzacţii. Fiecare dintre cei n indivizi $\boldsymbol{t_i}$ conţine articole alese din A, fiecare din cele p variabile Y^j este o variabilă cu valori binare care precizează pentru fiecare articol a_j faptul că acesta a fost ales sau nu în tranzacţia $\boldsymbol{t_i}$.

O submulţime P de articole din A, $P \subseteq A$, poartă numele de *itemset*. Dacă toate articolele conţinute într-un itemset P sunt conţinute şi în tranzacţia $\boldsymbol{t_i}$, $P \cap \boldsymbol{t_i} = P$**,** se spune că tranzacţia $\boldsymbol{t_i}$ *conţine* itemsetul P**,** $\boldsymbol{t_i} \supseteq P$. Numărul $\sigma(P) = \left| \{\boldsymbol{t_i} \mid \boldsymbol{t_i} \supseteq P, \boldsymbol{t_i} \in \boldsymbol{T}\} \right|$, al tranzacţiilor $\boldsymbol{t_i}$ ce conţin itemsetul P, se numeşte *suportul* (sau *susţinerea*) lui P.

O *regulă de asociere* între două itemseturi, P şi Q, este o expresie formală de tip implicaţie adică de forma: „ $P \rightarrow Q$ ” unde $P \cap Q = \varnothing$. *Puterea* unei reguli de asociere poate fi determinată pe baza a două metrici:

- $s(P \rightarrow Q) = \sigma(P \cup Q) / n$, numită *suport* (sau *susţinere*) şi
- $c(P \rightarrow Q) = \sigma(P \cup Q) / \sigma(P)$, numită *confidenţă* (sau *încredere*).

Suportul s exprimă măsura în care regula „ $P \rightarrow Q$ ” se aplică în mulţimea de observaţii disponibile, iar confidenţa c măsoară cât de frecvent articolele din Q apar în tranzacţiile care conţin P.

Cu aceste considerente problema analizei asocierilor se formulează astfel: *„Fiind dată o mulţime de tranzacţii,* $\boldsymbol{T}$*, să se găsească toate regulile „* $P \rightarrow Q$ *” care au suportul* $s \geq sup_{min}$ *şi confidenţa* $c \geq conf_{min}$ *, unde* sup_{min} *şi* $conf_{min}$ *sunt limite dorite de utilizator pentru* s *şi respectiv* c*”.*

Pattern-uri secvenţiale. Fie $A = \{1, \ldots, j, \ldots, p\}$ o mulţime de articole, fie $C = \{ c_1, \ldots, c_q\}$ mulţimea clienţilor şi fie $\boldsymbol{T} = \{\boldsymbol{t_1}, \ldots, \boldsymbol{t_i}, \ldots, \boldsymbol{t_n}\}$ mulţimea tranzacţiilor.

Fiecare din cele n tranzacţii $\boldsymbol{t_i}$ conţine p câmpuri pentru articole, un câmp , *id-client*, pentru identificarea clientului precum şi un câmp, *id-tr-time*, pentru precizarea momentului tranzacţiei; pentru orice client (id-client) există cel mult o tranzacţie la un moment dat (id-tr-time). Fiecare din cele p variabile Y^j este o variabilă cu valori binare şi precizează pentru fiecare articol j faptul că acesta a fost ales sau nu în tranzacţia $\boldsymbol{t_i}$, $\boldsymbol{t_i} \in \boldsymbol{T}$.

Orice submulţime $I \subseteq A$**,** de articole din A, poartă numele de *itemset.* O mulţime ordonată de itemseturi $S = \langle I_1, I_2, \ldots, I_s \rangle$ formează o *secvenţă*. Secvenţa $S = \langle I_1, I_2, \ldots, I_s \rangle$ *este conţinută* în secvenţa $R = \langle J_1, J_2, \ldots, J_r \rangle$, $S \subset R$, dacă există indicii $k_1 < k_2 < \ldots < k_s$ astfel încât $I_1 \subseteq J_{k1}$, $I_2 \subseteq J_{k2}$, ..., $I_s \subseteq J_{ks}$. Într-o mulţime de secvenţe Σ o secvenţă S^M este *maximală* dacă S^M nu este conţinută în nicio altă secvenţă din Σ: $(\forall) S \in \Sigma$, $S^M \not\subset S$.

Pentru o tranzactie $\boldsymbol{t_i} \in \boldsymbol{T}$, *itemset*($\boldsymbol{t_i}$) reprezintă itemsetul care conţine toate articolele alese în tranzacţia respectivă. Fie $c \in C$ şi fie $\boldsymbol{T^c} \subset \boldsymbol{T}$, $\boldsymbol{T^c} = \{ \boldsymbol{t^c_1}, \boldsymbol{t^c_2}, \ldots, \boldsymbol{t^c_{nc}}\}$ mulţimea tranzacţiior clientului c, ordonate crescător în timp (după id-tr-time). În aceste condiţii, se numeşte *secvenţă-client* secvenţa: $S(c) = \langle$ itemset($\boldsymbol{t^c_1}$), itemset($\boldsymbol{t^c_2}$), ..., itemset($\boldsymbol{t^c_{nc}}$) $\rangle$.

Se spune că un client oarecare, $c \in C$, *susţine* (sau *susţine*) secvenţa S, dacă S este conţinută în secvenţa-client a clientului c, $S \subset S(c)$. *Suportul unei secvenţe* S reprezintă fracţiunea clienţilor care susţin pe S: $s(S) = |\{c \mid c \in C, S \subset S(c)\}| \ / \ |C|$.

Problema descoperirii de pattern-uri secvențiale revine la a descoperi secvențele maximale, din mulțimea tuturor secvențelor Σ, care au un anumit suport minimal $\sup_{min}$ *, specificat de utilizator* : $S^M \in \{ S \mid s(S) \geq \sup_{min}\}$.

Orice astfel de secvență maximală reprezintă un *pattern secvențial*. O secvență care îndeplineşte condiția de suport minimal este numită *secvență mare*.

Analiza grupurilor. Se dispune de observații asupra a p variabile Y^j măsurate pe n indivizi. Fie $X = \{x_1, ..., x_i, ..., x_n\}$ mulțimea celor n indivizi caracterizați de cele p variabile; se presupune că spațiul $\Re^p$, ce conține pe X, este dotat cu o distanță (euclidiană sau χ^2) sau cu o similaritate. Se doreşte *partiționarea* mulțimii X în k *submulțimi* (*clase* sau *clustere*), unde k este cunoscut à priori, astfel încât clasele C_l, $l = 1 \div k$, obținute să fie cât mai *omogene*.

Fie $g_1, g_2, ..., g_i, ..., g_k$ *centrele de greutate* ale celor k clase dorite : *inerția clasei* C_l este $I_l = \sum_{x_i \in C_l} p_j d^2(x_i, g_i)$, cu p_j s-a notat ponderea individului x_i ; *inerția intraclase* este $I_W = \sum_{l=1}^{k} P_l I_l$, unde P_l este ponderea clasei l (numărul de indivizi); *inerția interclase* este $I_B = \sum_{l=1}^{k} P_l d^2(g_l, g)$, unde g este centrul de greutate al mulțimi X de n indivizi; *inerția totală* a lui X, este $I = I_W + I_B$, principiul lui König-Huygens. O clasă este cu atât mai omogenă cu cât inerția mulțimii de puncte ce o alcătuieşte este mai mică.

Un criteriu de partiționare pentru a determina, în medie, clase omogene, constă în a căuta acea partiție în k clase pentru care inerția intraclase este minimă, deci inerția interclase este maximă.

Clasificare. Se dispune de observații privind p variabile cantitative Y^j şi o variabilă nominală având q modalități Y, măsurate pe n indivizi. Cei n indivizi sunt împărțiți în q clase presupuse disjuncte, definite à priori de variabila nominală Y şi se cunoaşte afectarea fiecărui individ la o clasă.

Fie $X = \{x_1, ..., x_i, ..., x_n\}$ mulțimea celor n indivizi caracterizați de cele p variabile şi fie y vectorul n-dimensional cu componentele y_i ($i = 1 \div n$, $y_i \in \{1, ..., q\}$) reprezentând numărul clasei din care face parte individul x_i. Problema de clasificare (sau de clasare), respectiv problema afectării unui individ suplimentar x_s, caracterizat prin cele p variabile, la una dintre cele q clase poate fi formulată după cum urmează:

Pe baza datelor disponibile $\{(x_1, y_1), ..., (x_i, y_i), ..., (x_n, y_n)\}$, *numit eşantion de învățare*, *să se definească o regulă* (*sau un clasificator*) $\varphi(\bullet)$, *astfel încât* $\varphi(\bullet)$ *să poată fi evaluat pentru orice individ x, nu doar pentru cei incluşi în datele de învățare iar clasa atribuită oricărui individ nou* x_s, $\hat{y} = \varphi(x_s)$, *să fie cât mai apropiată posibil de clasa reală y*.

Pentru datele x_i din eşantionul de învățare, clasele reale, y_i, sunt cunoscute, dar nu vor coincide în mod necesar cu aproximările lor $\hat{y}_i = \varphi(x_i)$. Pentru indivizii x_s noi, clasele reale y_s nu sunt cunoscute, dar principala țintă a procedurii de clasificare este ca aproximarea $\hat{y}_s = \varphi(x_s) \approx y_s$ să fie cea mai bună posibilă, motiv pentru care calitatea aceastei aproximări trebuie să fie judecată pe baza proprietăților statistice sau probabilistice ale întregii populații din care viitorii indivizi vor fi preluați.

Mulțimi „rough". Se dispune de un *sistem de informații* adică de un cuplu de mulțimi finite şi nevide $I = (U, V)$ unde: $U = \{i \mid i = 1 \div n\}$ este o mulțime de *obiecte* (sau de *indivizi*) numită *univers*; $V = \{V^j \mid V^j : U \to \nabla_j, j = 1 \div p\}$ este o mulțime de *caracteristici* (sau de *variabile*) astfel încât sistemul asignează caracterisicii j a individului i valoarea x_{ij} , $x_{ij} = V^j(i)$, din domeniul ∇_j al funcției V^j.

Doi indivizi $i, \ell \in U$, $i \neq \ell$, pentru care $x_{ij} = x_{\ell j}$ se numesc *indiscernabili* în raport cu variabila V^j.

Fie acum W o submulțime de variabile, $W \subseteq V$ şi fie $ind(W) \subseteq U \times U$ relația indusă de W *pe* indivizii din U, unde $ind(W) = \{(i, \ell) \mid i, \ell \in U, x_{ij} = x_{\ell j} \ (\forall) V^j \in W\}$ şi este o relație de echivalență pe U; $(i, \ell) \in ind(W) \Leftrightarrow$ indivizii i, ℓ sunt indiscernabili în raport cu variabilele din W. Relația $ind(W)$, numită şi relație de *W-indiscernabilitate*, partiționează mulțimea U în clase de echivalență notate prin $[i]_W$, $U_{ind(W)} = \cup [i]_W$.

$V = P \cup D$ unde P reprezintă mulțimea variabilelor *de predicție* (sau *explicative*), iar D mulțimea variabilelor *de decizie* (sau *de explicat*). Se numește *matrice de discernabilitate* matricea $M \in \mathcal{M}_{n \times n}(\Re)$ ale cărei elemente

$$m_{i\ell} = \{p \in P \mid [V^p(i) \neq V^p(\ell)] \wedge [(V^d(i) \neq V^d(\ell), (\forall) V^d \in D]\}$$

reprezintă liste de variabile de predicție care plasează indivizii i, ℓ în clase diferite ale partiției $U_{ind(D)}$.

Fie $Z \subseteq U$ și $W \subseteq P$**.** Se dorește ca mulțimea țintă Z să fie descrisă cu ajutorul variabilelor din W. Descrierea lui Z nu poate fi precisă deoarece pentru anumite obiecte din U, indiscernabile în raport cu variabilele din W, nu se știe dacă pot fi incluse sau nu în Z și atunci descrierea lui Z va fi aproximativă.

Se numește *aproximare W-inferioară* a lui Z mulțimea indivizilor care pot fi clasați ca membri *siguri* ai lui Z: $W_{inf}Z = \{ i \in U \mid [i]_W \subseteq Z \}$. Se numește *aproximare W-superioară* a lui Z mulțimea indivizilor ce pot fi clasați membri *posibili* ai lui Z: $W^{sup}Z = \{ i \in U \mid [i]_W \cap Z \neq \varnothing \}$. Se numește *acuratețe* a aproximării raportul: $\alpha_W(Z) = |W_{inf}Z| / |W^{sup}Z|$. Se numește *regiune de frontieră* a lui Z mulțimea indivizilor care nu pot fi clasați cu certitudine nici în Z nici în afara lui Z: $W_fZ = W^{sup}Z - W_{inf}Z$**.**

Dacă $W_f(Z) \neq \varnothing$ **,** mulțimea Z se numește *mulțime rough* sau *W-rough* (pe baza cunoștințelor din W), iar dacă $W_fZ = \varnothing$**,** mulțimea Z se numește *mulțime crisp* sau *W-definibilă* (pe baza cunoștințelor din W).

Pentru un sistem de informații (U, V) *în care se evidențiază situații de indiscernabilitate, abordarea bazată pe mulțimi rough permite clasări ale indivizilor din U, pe baza partițiilor induse de variabile din V, prin determinarea unor submulțimi de aproximare inferioară, de aproximare superioară și de frontieră.*

„Link mining". Mulțimea de obiecte (sau de indivizi) observată este mulțimea paginilor Web. Fie $G = (P, L)$ un graf orientat asociat spațiului Web, unde: P, mulțimea nodurilor, reprezintă mulțimea paginilor Web și L, mulțimea arcelor orientate, reprezintă mulțimea (hyper)link-urilor. Se numește *in-link* (*out-link*) al unei pagini i orice link care indică (din) pagina i din (către) alte pagini.

Un link de la pagina ℓ la pagina i este considerat ca un transfer implicit de autoritate către pagina i și din acest punct de vedere o pagină cu mai multe in-link-uri este considerată a fi *de calitate* mai înaltă (sau cu un *scor de calitate* mai mare) decât o altă pagină cu mai puține in-link-uri astfel încât, din punctul de vedere al calității paginile pot fi pagini *de* (înaltă) *calitate* și *pagini comune*.

Din punctul de vedere al timpului paginile pot fi *pagini vechi* și *pagini noi*, adăugate recent. Paginile vechi, dacă sunt de calitate și sunt actualizate rămân de calitate, în caz contrar devin, în timp, comune, iar dacă sunt comune și sunt actualizate pot deveni, în timp, de calitate, în caz contrar rămânând comune. Paginile noi, similar cu articolele științifice noi (preferate de către cercetători), deși pot fi de calitate, fiind publicate recent scorul acestora (numărul de in-link-uri sau de citări) este de așteptat să fie foarte mic sau chiar nul.

Fiind dată o cerere de căutare, a unui utilizator, principala sarcină a motoarelor de căutare este de a găsi paginile relevante, de cea mai înaltă calitate, care satisfac nevoia de informare a utilizatorului.

4. Obiective și strategii data mining

Ceea ce se exploatează prin data mining sunt colecții de date disponibile, de volum mare sau foarte mare, provenite din surse interne ale organizației care au fost constituite, inclusiv ca structură, în perspectiva altor finalități, și la care se adaugă date provenite din diverse alte surse externe organizației [5, 6, 8, 10]. Utilizarea data mining presupune:

- identificarea oportunității acesteia și a datelor pe care se poate baza explorarea;

- extragerea informațiilor din colecțiile / depozitele de date existente şi prelucrarea acestora prin tehnici adecvate de data mining;
- adoptarea de decizii pe baza rezultatelor obținute şi întreprinderea de acțiuni;
- măsurarea rezultatelor concrete pentru a identifica şi alte modalități de exploatare a datelor disponibile.

Un prim demers, de multe ori plictisitor dar inevitabil, constă în efectuarea unei *explorări* a acestor date: alura distribuțiilor, prezența datelor atipice, corelații şi coerență, transformări eventuale ale datelor; clasificare.

Demersul descriptiv şi exploratoriu permite realizarea de rezumate şi grafice mai mult sau mai puțin elaborate, descrierea mulțimilor de date şi stabilirea de relații între variabile, fără a acorda un rol privilegiat vreunei variabile. Demersul exploratoriu se sprijină, în mod esențial, pe noțiuni elementare (medie şi dispersie), pe reprezentări grafice şi pe tehnici descriptive multidimensionale. Metodele exploratorii caută subspațiile de reprezentare (factoriale) de dimensiuni mici, care aproximează cel mai bine norii de puncte-indivizi sau de puncte-variabile, astfel încât vecinătățile măsurate în aceste spații să reflecte cât mai exact proximitățile reale.

În demersul descriptiv şi exploratoriu *obiectivele principale* urmărite sunt:

- *explorare multidimensională*, bazată cel mai frecvent pe metode precum *analiza în componente principale*, *analiza factorială discriminantă*, *analiza corespondențelor simple*, *analiza corespondențelor multiple* şi *analiza canonică*.
- *clasificare*, utilizând cel mai adesea metode precum *clasificarea ascendentă ierarhică*, *metoda norilor dinamici* sau o *metodă mixtă*.

Un al doilea demers îl constituie modelarea în scopul predicției unei (unor) variabile țintă prin variabilele explicative utilizând instrumente de modelare (sau de învățare).

Demersul inferențial şi *confirmatoriu* permite validarea (sau infirmarea), pornind de la teste statistice sau modele probabiliste, a ipotezelor formulate à priori (adică urmare a unui demers exploratoriu) şi extrapolarea acestora de la nivelul eşantionului la cel al unei populații mai mari. Demersul confirmatoriu face apel, în special, la metodele numite explicative şi previzionale destinate să explice apoi să prevadă, urmând anumite reguli de decizie, o variabilă privilegiată cu ajutorul uneia sau mai multor variabile explicative.

În demersul inferențial şi confirmatoriu *obiectivul principal* urmărit îl constituie *modelarea*/*discriminarea* respectiv deducerea unui model de previziune pentru variabila (variabilele) țintă. Metodele cele mai frecvent utilizate în atingerea acestui obiectiv sunt: *modelul liniar general*, *analiza discriminantă*, *rețelele neuronale*, *mașinile cu suport vectorial*, *arborii de clasificare şi de regresie*, *agregarea modelelor* ("*Bagging"*, "*Boosting"*, "*Random Forest"*).

Demersurile sunt complementare, explorarea şi descrierea trebuind, în general, să preceadă etapele explicative şi predictive [1, 2, 3]. O explorare preliminară este adesea utilă pentru a avea o primă idee despre natura legăturilor între variabile şi pentru a trata cu prudență variabilele corelate, şi deci redundante, ce riscă să încarce inutil modelul. *Succesiunea acestor două demersuri, explorare şi apoi învățare, constituie fundamentul utilizării data mining*. Spre deosebire de abordarea statistică tradițională, în care observarea datelor este integrată în metodologie (planificarea experimentului), în data mining datele sunt *prealabile* analizei. Pentru a se oferi şanse mai favorabile de succes unui proces data mining este evident că preocupările legate de definirea obiectivelor şi de analiză a datelor ar trebui să intervină cât mai devreme posibil.

Strategiile uzuale pentru data mining constau din înlănțuirea a patru etape majore:

- *extracție*; extragerea datelor, eventual prin sondaj.
- *explorare*; studiul distribuțiilor, transformare, recodificarea eventuală a variabilelor cantitative, regruparea modalităților variabilelor calitative, eliminarea anumitor variabile, selecționarea acelora cel mai strâns legate de variabila țintă, completarea datelor lipsă, cercetarea eventualelor relații neliniare.

- *analiză*;
    - *clasificare*: caracterizarea claselor prin variabilele inițiale cu ajutorul instrumentelor de discriminare,
    - *modelare / discriminare*: extracția unui eşantion de test, estimarea, optimizarea modelelor pentru fiecare din metodele utilizabile (validare încrucişată), compararea performanțelor,
- *exploatare*; odată ce o metodă asociată cu un model sunt considerate ca fiind bine alese întregul eşantion este regrupat pentru a face o ultimă estimare a modelului, exploatarea modelului şi difuzarea rezultatelor.

Există produse informatice de data mining putând funcționa pe arhitecturi de tip client-server menite să exploateze volume foarte mari de date, cu paletă largă de tehnici atât în variantă statistică cât şi în variantă data mining. Există, de asemenea, numeroase produse informatice de data mining realizate pentru PC-uri, simplu de instalat, nu foarte scumpe, cu algoritmi de bună calitate, conviviale şi suficiente pentru IMM (prelucrând zeci şi chiar sute de mii de linii) şi care oferă în general una sau două tehnici de data mining.

Sistemele software de data mining sunt, în general, capabile să asigure:

- *algoritmi*: de *clasare* (analiză discriminantă liniară, regresie logistică binară sau politomică, model liniar generalizat, regresie logistică PLS, arbori de decizie, rețele neuronale, k-vecini cei mai apropiați); de *predicție* (regresie liniară, model liniar general, regresie robustă, regresie neliniară, regresie PLS, arbori de decizie, rețele neuronale, k-vecini cei mai apropiați); de *clasificare* (centre mobile, nori dinamici, k-medii, clasificare ierarhică, metoda mixtă, rețele Kohonen); de *analiză a seriilor temporale*; de *analiză a fiabilității* (supraviețuirii); de *detecție a asocierilor*.
- *funcții de pregătire a datelor*: de manevrare fişiere (fuziune, agregare, transpoziție); de vizualizare indivizi, colorare conform unui criteriu dat; de detectare, filtrare şi tratare extreme; de analiză şi tratare valori lipsă; de transformare a variabilelor (recodificare, standardizare, normalizare automată, discretizare); de creare de noi variabile (funcții logice, şiruri, statistici, funcții matematice); de selecție a discretizărilor, interacțiunilor şi variabilelor celor mai explicative.
- *funcții de prelucrări statistice*: determinarea caracteristicilor de tendință centrală, de dispersie, de formă; teste statistice de medie, de varianță, de distribuție, de independență, de heteroscedasticitate, de multicoliniaritate.
- *funcții de eşantionare şi de partiționare a datelor*: crearea de eşantioane de învățare, de test şi de validare (eşantionarea stratificată trebuie să fie posibilă); „Bootstrap", „jackknife" (validare încrucişată).
- *funcții de analiză exploratorie a datelor*.
- *limbaje evoluate de programare* (bazate pe macro-instrucțiuni).
- *facilități de prezentare a rezultatelor*: vizualizare rezultate, manipulare tabele, biblioteci de grafice (2D, 3D, interactiv), navigare în arbori de decizie, afişare curbe, indice Gini, încorporare rezultate în diverse rapoarte.
- *facilități de gestiune a metadatelor* (definirea uniformă a tuturor variabilelor şi a grupurilor de variabile).
- *platforme suport* (Windows, Unix, Sun, IBM-MVS); formate de intrare/ieşire ale datelor gestionate (tabele Oracle, Sybase, DB2, SAS, Excel); volume de date care pot fi rezonabil tratate.
- *putere de calcul* (arhitecturi client server, calcule pe server vizualizarea rezultatelor pe client; algoritmi paraleli); execuția în mod interactiv sau diferit; portabilitatea modelelor construite (C, XML, Java, SQL).

În cadrul procesului decizional, mai larg, *procesul data mining* se desfăşoară ca o succesiune de faze:

- extragerea datelor, cu sau fără eşantionare, recurgând la tehnici de sondaj aplicate sau aplicabile bazelor de date;
- explorarea datelor pentru detectarea valorilor aberante sau doar atipice, a incoerenţelor, pentru studiul de distribuţiilor, structurilor de corelaţie, pentru căutarea tipologiilor, pentru transformarea datelor;
- partiţionarea aleatoare a eşantionului (învăţare, validare, testare), în funcţie de mărimea acestuia şi de tehnicile care vor fi utilizate, pentru a estima o eroare de predicţie în vederea alegerii modelului, a alegerii şi certificării metodei;
- pentru fiecare din metodele luate în consideraţie: estimarea modelului pentru o valoare dată unui parametru de *complexitate* (numărul de variabile, de vecini, de frunze, de neuroni, durata de învăţare, etc.) şi optimizarea acestui parametru;
- compararea modelelor optimale obţinute (câte unul pentru fiecare metodă) prin estimarea erorii de previziune;
- iterarea eventuală a etapelor precedente, în cazul în care eşantionul de test este prea mic. Partiţionări aleatoare succesive ale eşantionului pentru medierea pe mai multe cazuri a estimării finale a erorii de predicţie şi asigurarea robusteţii modelului obţinut;
- alegerea metodei adoptate, pe baza capacităţilor sale de predicţie, a robusteţii sale dar şi, eventual, a interpretabilităţii modelului obţinut.

5. Concluzii

O practică bună de data mining necesită din partea asistenţilor decizionali să ştie să articuleze toate metodele [1, 2, 4, 5] sarcină care nu poate fi îndeplinită decât cu condiţia de a avea foarte bine clarificate obiectivele studiului.

Pe de o parte, multe metode urmăresc aceleaşi obiective predictive. În cazurile fericite, când datele sunt bine structurate, metodele furnizează rezultate foarte asemănătoare. În celelalte cazuri o anumită metodă poate să se dovedească mai eficace, fie datorită mărimii eşantionului, fie că geometric este mai bine adaptată topologiei grupurilor de discriminat, fie datorită mai bunei interacţiuni cu tipurile de variabile. Astfel, în multe situaţii, poate fi esenţială şi eficace o decupare în clase de variabile predictive cantitative pentru a aborda în mod restrâns o versiune neliniară a modelului prin combinarea variabilelor auxiliare. Acest aspect poate fi important de exemplu în cazul regresiei logistice sau perceptronului, dar este inutil în cazul arborilor de decizie care integrează acest decupaj în clase în chiar construcţia modelelor (singurele optimale).

Pe de altă parte, metodele nu prezintă toate aceleaşi facilităţi de interpretare. Nu există o cea mai bună alegere à priori. Numai experienţa şi un protocol de test atent construit permit determinarea acesteia. Este şi motivul pentru care sistemele software generaliste nu fac o alegere şi oferă aceste metode în paralel pentru a se adapta mai bine la date, la deprinderile fiecărui utilizator (client potenţial) şi chiar şi "modei".

În fazele exploratorii pot fi găsite relaţii care aparent au semnificaţii importante, valabile în interiorul setului de testare, dar care s-ar putea să fie fără nici o semnificaţie statistică întro populaţie mai largă („*data dredging"*, „*data fishing"*, „*data snooping"*).

În fazele de modelare, o supraparametrizare sau o supraajustare a modelului poate explica perfect datele fără ca rezultatele să fie totuşi extrapolabile sau generalizabile la alte date decât cele studiate. Rezultatele previziunii pot fi deci viciate de o importantă eroare relativă legată de varianţa estimaţiilor parametrilor. Problema este de a găsi un compromis bun între bias-ul unui model mai mult sau mai puţin fals şi varianţa estimatorilor.

Obiectivul esenţial rămâne „căutarea sensului" în vederea facilitării luărilor de decizie,

prezervând fiabilitatea. Prezența sau controlul unei expertize statistice rămâne inevitabilă pentru că necunoaşterea limitelor şi capcanelor metodelor utilizate poate conduce la aberații de natură să discrediteze demersul, făcând caduce investițiile consimțite.

BIBLIOGRAFIE

1. **BACCINI, A.; BESSE, P.:** Data mining / Exploration Statistique. Toulouse: INSA, 2010, 111 p.
2. **BESSE, P.:** Apprentissage Statistique & Data mining. Toulouse: INSA, 2009, 124 p.
3. **ENĂCHESCU, D.:** Data Mining - metode şi aplicații. Bucureşti: Editura Academiei Române, 2009, 277 p.
4. **FILIP, F. G.:** Decizie asistată de calculator: decizii, decidenți – metode de bază şi instrumente informatice asociate, Ed. a 2-a, rev. Bucureşti: Editura Tehnică, 2005, 376 p.
5. **FILIP, F. G.:** Sisteme suport pentru decizii, Ed. a 2-a, rev. Bucureşti: Ed. Tehnică, 2007, 364 p.
6. **GORUNESCU, F.:** Data Mining, Concepts, Models and Techniques, Springer- Heidelberg, series Intelligent Systems Reference Library, 2011, 372 p.
7. **HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J.:** The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition. Springer-Verlag, Springer Series in Statistics, 2008, 763 p.
8. **MĂRGINEAN, N.:** Sisteme inteligente pentru asistarea deciziilor. Editura Risoprint, Cluj-Napoca, 2006, 239 p.
9. **PENG, Y.; KOU, G.; SHI, Y.; CHEN, Z.:** A descriptive framework for the field of data mining and knowledge discovery. International Journal of Information Technology & Decision Making, Vol. 7, No. 4, 2008, pp. 639-682.
10. **TAN, P-N.; STEINBACH, M.; KUMAR, V.:** Introduction to Data Mining. Addison-Wesley, 2006, 769 p.
11. **TUFFERY, S.:** Data mining et statistique décisionnelle, 3ème Edition. Editions TECHNIP, 2010, 705 p.
12. **WU, X.; KUMAR, V. (ED.):** The Top Ten Algorithms in Data Mining. Chapman & Hall / CRC DMKD Series, 2009, 232 p.
13. **YU, P-S.; HAH, J.; FALOUSTOS, C. (ED.):** Link Mining: Models, Algorithms, and Applications. Springer, 2010, 586 p.