

EXPLORAREA DATELOR ȘI DESCOPERIREA CUNOȘTIȚELOR - PROBLEME, OBIECTIVE ȘI STRATEGII -

Cornel Lepădatu (cornel_lepadatu@biblacad.ro)
Biblioteca Academiei Române, Academia Română
București

Rezumat

Explorarea datelor și descoperirea cunoștințelor, „data mining”, este un ansamblu de metode și algoritmi destinat explorării și analizei unor, adesea, mari volume de date în vederea deducerii, din aceste date, a unor reguli, a unor asocieri, a unor tendințe 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 și lipsit de o abordare integrată, situație care provoacă dificultăți în procesele de predare, de învățare, de cercetare precum și 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 și formula problemele.

Articolul se concentrează asupra unor aspecte importante privind natura și calitatea datelor utilizate în aplicarea data mining, privind metodele cele mai frecvent utilizate, alegerea principalelor obiective, formularea și 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

DATA MINING AND KNOWLEDGE DISCOVERY – ISSUES, OBJECTIVES AND STRATEGIES –

Abstract

Data mining and knowledge discovery denote a set of methods and algorithms for exploration and analysis of (often) large volumes of data for 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 the 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 și sute de mii de înregistrări privitoare la cumpărările, vânzările, încasările și plățile efectuate;
- societățile de telefonie mobilă posedă date privitoare la fiecare convorbire efectuată de abonații lor, incluzând data, momentul și 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 și 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 și 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 și tehnicile informatice, tot mai evoluat, au contribuit de-a lungul timpului la amplificarea capacității de memorare și 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 și simplu în virtutea nevoii de arhivare;
- datele acumulate conțin informații și cunoștințe „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 și înțelegerea fenomenelor care au guvernat funcționarea surselor de date;
- creșterea permanentă a concurenței, exigențele din ce în ce mai mari ale pieței au determinat organizațiile să devină conștiente de potențialul pe care aceste arhive de date îl reprezintă.

„Informația nu lipsește, ceea ce lipsește este timpul managerului de a considera toate informațiile care sunt disponibile” semnală 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 și tipul acestor informații este foarte diversificat:

- web-ul este un exemplu, foarte prezent astăzi, de spațiu care regroupează date foarte numeroase, diverse și 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 cunoaștere utilă:

- pe web, este vorba de a înțelege mai bine conținutul paginilor web și cererile utilizatorilor pentru a le furniza informația țintă cea mai pertinentă posibilă și 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 privește datele provenite din procesele de producție, există un mare interes de a extrage din ele un maximum de cunoștințe 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 și pentru a acționa în consecință pentru populațiile vizate.

2. Tehnologia data mining

Preocupările privind descoperirea de noi cunoștințe 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 științele sociale și în inteligența artificială.

Data mining nu este nici noutate tehnologică nici științifică, metodele și tehnicile utilizate sunt relativ vechi. Noutatea a constat în integrarea acestora în procesarea industrială a informației. Dezvoltarea în decursul timpului a diverselor concepte, metode și 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 corelație*, Karl Pearson; 1900, *distribuția χ^2* , Karl Pearson; 1930, *analiza factorială*, Hotteling; 1936, *analiza discriminantă*, Fisher și Mahalanobis; 1941, *analiza factorială a corespondențelor*, Guttman; 1943, *rețele neuronale*, Mc Culloch și Pitts; 1944, *regresia logistică*, Joseph Berkson; 1958, *perceptronul*, Rosenblatt; 1962, *analiza datelor*, J.-P. Benzécri; 1964, *arbore de decizie AID*, J.P.Sonquist și J.-A.Morgan; 1965, *metoda centrelor mobile*, E. W. Forgy; 1967, *metoda celor k-medii*, Mac Queen; 1972, *modelul liniar generalizat*, Nelder și Wedderburn.
- *analiza datelor* (sau istorie): 1975, *algoritmi genetici*, Holland; 1975, *metoda de clasare DISQUAL*, Gilbert Saporta; 1977, *analiza exploratorie a datelor*, Tukey; 1980, *rețele bayesiene*, Pearl; 1980, *arbore de decizie CHAID*, KASS; 1983, *regresie PLS* (Partial Least Squares), Herman și Svante Wold; 1984, *arbore CART*, Breiman, Friedman, Olshen, Stone; 1986, *perceptron multistrat*, Rumelhart și McClelland; 1989, *rețele (auto-adaptative)*, T. Kohonen.
- *explorarea datelor și descoperirea cunoștințelor*: **1990** (aproximativ), apariția conceptului de **data mining**; 1993, *arbore C4.5*, J. Ross Quinlan; 1996, *bagging* (Breiman) și *boosting* (Freund-Shapire); 1998, *mașini 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 obținute cu o procedură specială (eșantionare, planificare experiment) ; ipoteze tari privind legile

statistice urmate ; modelele provin din teorie și sunt confruntate cu datele ; metode probabiliste și statistice ; utilizare în laborator.

- *analiza datelor* : câteva zeci de mii de indivizi ; câteva zeci de variabile ; construirea de tabele "indivizi \times variabile" ; importanță pentru calcul și reprezentare vizuală.

- *data mining* : mai multe milioane de indivizi ; mai multe sute de variabile ; numeroase variabile nenumerice, uneori textuale ; date obținute anterior studiului și adesea în alte scopuri ; date imperfecte, erori de obținere sau de codificare, valori lipsă sau aberante ; populație constant evolutivă (dificil de eșantionat) ; necesitatea unor calcule rapide, uneori în timp real ; nu se caută întotdeauna optimul matematic ci modelul cel mai ușor de înțeles (sau de asimilat) de către utilizatori ; ipoteze slabe privind legile statistice urmate ; modelele sunt obținute din date din care se deduc și elemente teoretice ; metode statistice, de inteligență artificială și de teoria învățării ; utilizare în organizații.

3. Probleme data mining

Tehnologia data mining permite descoperirea de pattern-uri structurale din date utilizând algoritmi suficient de robuști atât pentru a prelucra date imperfecte, corelate stohastic, cât și pentru a extrage corelații, uneori imprecise, și reguli utilizabile ulterior în predicția, explicarea și înțelegerea evoluției structurii datelor analizate.

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

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

Analiza asocierilor. Fie $A = \{ a_1, \dots, a_j, \dots, a_p \}$ o mulțime de articole și fie $T = \{ t_1, \dots, t_i, \dots, t_n \}$ o mulțime de tranzacții. Fiecare dintre cei n indivizi t_i conține 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 tranzacția t_i .

O submulțime P de articole din A , $P \subseteq A$, poartă numele de *itemset*. Dacă toate articolele conținute într-un itemset P sunt conținute și în tranzacția t_i , $P \cap t_i = P$, se spune că tranzacția t_i conține itemsetul P , $t_i \supseteq P$. Numărul tranzacțiilor t_i ce conțin itemsetul P , $\sigma(P) = | \{ t_i \mid t_i \supseteq P, t_i \in T \} |$, se numește *suportul* (sau *susținerea*) lui P .

O *regulă de asociere* între două itemseturi, P și Q , este o expresie formală de tip implicație adică de forma: „ $P \rightarrow Q$ ” unde $P \cap Q = \emptyset$. *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 *susținere*) și
- $c(P \rightarrow Q) = \sigma(P \cup Q) / \sigma(P)$, numită *confidență* (sau *încredere*).

Supportul s exprimă măsura în care regula „ $P \rightarrow Q$ ” se aplică în mulțimea de observații disponibile iar confidența c măsoară cât de frecvent articolele din Q apar în tranzacțiile care conțin P .

Cu aceste considerente problema analizei asocierilor se formulează astfel: „Fiind dată o mulțime de tranzacții, T , să se găsească toate regulile „ $P \rightarrow Q$ ” care au suportul $s \geq \text{sup}_{\min}$ și confidența $c \geq \text{conf}_{\min}$, unde sup_{\min} și conf_{\min} sunt limite dorite de utilizator pentru s și respectiv c ”.

Pattern-uri secvențiale. Fie $A = \{1, \dots, j, \dots, p\}$ o mulțime de articole, fie $C = \{c_1, \dots, c_l, \dots, c_q\}$ mulțimea clienților și fie $T = \{t_1, \dots, t_i, \dots, t_n\}$ mulțimea tranzacțiilor.

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

Orice submulțime $I \subseteq A$, de articole din A , poartă numele de *itemset*. O mulțime ordonată de itemseturi $S = \langle I_1, I_2, \dots, I_s \rangle$ formează o *secvență*. Secvența $S = \langle I_1, I_2, \dots, I_s \rangle$ este conținută în secvența $R = \langle J_1, J_2, \dots, J_r \rangle$, $S \subset R$, dacă există indicii $k_1 < k_2 < \dots < k_s$ astfel încât $I_1 \subseteq J_{k_1}$, $I_2 \subseteq J_{k_2}$, ..., $I_s \subseteq J_{k_s}$. Într-o mulțime de secvențe Σ o secvență S^M este *maximală* dacă S^M nu este conținută în nicio altă secvență din Σ : $(\forall) S \in \Sigma, S^M \not\subset S$.

Pentru o tranzacție $t_i \in T$, $itemset(t_i)$ reprezintă itemsetul care conține toate articolele alese în tranzacția respectivă. Fie $c \in C$ și fie $T^c \subset T$, $T^c = \{t^c_1, t^c_2, \dots, t^c_{nc}\}$ mulțimea tranzacțiilor clientului c , ordonate crescător în timp (după câmpul $id\text{-}tr\text{-}time$). În aceste condiții, se numește *secvență-client* secvența: $S(c) = \langle itemset(t^c_1), itemset(t^c_2), \dots, itemset(t^c_{nc}) \rangle$.

Se spune că un client oarecare, $c \in C$, *suportă* (sau *susține*) secvența S , dacă S este conținută în secvența-client a clientului c , $S \subset S(c)$. *Supportul unei secvențe* S reprezintă fracțiunea clienților care susțin 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: $S^M \in \{S \mid s(S) \geq \text{sup}_{\min}\}$.

Orice astfel de secvență maximală reprezintă un *pattern secvențial*. O secvență care îndeplinește 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 $X = \{x_1, \dots, x_i, \dots, x_n\}$ mulțimea celor n indivizi caracterizați de cele p variabile; se presupune că spațiul \mathfrak{R}^p , ce conține pe X , este dotat cu o distanță (euclidiană sau χ^2) sau cu o similaritate. Se dorește *partiționarea* mulțimii X în k submulțimi (clase sau *cluster*), unde k este cunoscut a priori, astfel încât clasele $C_i, 1 = 1 \div k$, obținute să fie cât mai *omogene*.

Fie $g_1, g_2, \dots, g_i, \dots, g_k$ centrele de greutate ale celor k clase dorite: *inerția clasei* C_1 este $I_1 = \sum_{x_i \in C_1} p_j d^2(x_i, g_1)$, cu p_j s-a notat ponderea individului x_i ; *inerția intraclase* este $I_W = \sum_{i=1}^k P_i I_i$, unde P_i reprezintă ponderea clasei i (numărul de indivizi); *inerția interclase* este $I_B = \sum_{i=1}^k P_i d^2(g_i, g)$, unde g este centrul de greutate al mulțimii X de n indivizi; *inerția totală* a mulțimii 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 alcătuiește 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 și o variabilă nominală având q modalități Y , măsurate pe n indivizi. Cei n indivizi sunt împărțiți în q clase presupuse disjuncte, definite a priori de variabila nominală Y și se cunoaște afectarea fiecărui individ la o clasă.

Fie $X = \{x_1, \dots, x_i, \dots, x_n\}$ mulțimea celor n indivizi caracterizați de cele p variabile și fie y vectorul n -dimensional cu componentele y_i ($i = 1 \div n$, $y_i \in \{1, \dots, q\}$) 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 $\{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$, numit eșantion 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 incluși în datele de învățare; 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 eșantionul 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 acestei 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 și nevide $I = (U, V)$ unde: $U = \{i \mid i = 1 \div n\}$ este o mulțime de obiecte (sau de indivizi) numită univers; $V = \{V^j \mid V^j : U \rightarrow \mathcal{V}_j, j = 1 \div p\}$ este o mulțime de caracteristici (sau de variabile) astfel încât sistemul asignează caracteristicii j a individului i valoarea x_{ij} din domeniul \mathcal{V}_j al funcției V^j , $x_{ij} = V^j(i)$.

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, $W \subseteq V$. Relația $ind(W) \subseteq U \times U$, unde $ind(W) = \{(i, \ell) \mid i, \ell \in U, x_{ij} = x_{\ell j} \ (\forall) V^j \in W\}$ este o relație de echivalență pe U și se numește *relație indusă de W pe indivizii din U* ; $(i, \ell) \in ind(W) \Leftrightarrow$ indivizii i, ℓ sunt indiscernabili în raport cu variabilele din W . Relația $ind(W)$, numită și 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 $M \in \mathcal{M}_{n \times n}(\mathfrak{R})$ 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 care pot fi clasați ca membri *posibili* ai lui Z : $W^{sup}Z = \{ i \in U \mid [i]_W \cap Z \neq \emptyset \}$. 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 $W_f(Z) \neq \emptyset$, mulțimea Z se numește *mulțime rough* sau *W-rough* (pe baza cunoștințelor din W), iar dacă $W_fZ = \emptyset$, 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 (întă) 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 și prelucrarea acestora prin tehnici adecvate de data mining;
- adoptarea de decizii pe baza rezultatelor obținute și întreprinderea de acțiuni;
- măsurarea rezultatelor concrete pentru a identifica și 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 și coerență, transformări eventuale ale datelor; clasificare.

Demersul descriptiv și exploratoriu permite realizarea de rezumate și grafice mai mult sau mai puțin elaborate, descrierea mulțimilor de date și stabilirea de relații între variabile, fără a acorda un rol privilegiat vreunei variabile. Demersul exploratoriu se sprijină, în mod esențial, pe noțiuni elementare (medie și dispersie), pe reprezentări grafice și 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 și 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* și *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 și confirmatoriu permite validarea (sau infirmarea), pornind de la teste statistice sau modele probabiliste, a ipotezelor formulate a priori (adică urmare a unui demers exploratoriu) și extrapolarea acestora de la nivelul eșantionului la cel al unei populații mai mari. Demersul confirmatoriu face apel, în special, la metodele numite explicative și 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 și 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 și de regresie*, *agregarea modelelor* (“*Bagging*”, “*Boosting*”, “*Random Forest*”).

Demersurile sunt complementare, explorarea și descrierea trebuind, în general, să preceadă etapele explicative și predictive [1, 2, 3]. O explorare preliminară este adesea utilă pentru a avea o primă idee despre natura legăturilor între variabile și pentru a trata cu prudență variabilele corelate, și deci redundante, ce riscă să încarce inutil modelul. *Sucesiunea acestor două demersuri, explorare și 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 *prelabile* analizei. Pentru a se oferi șanse mai favorabile de succes unui proces data mining este evident că preocupările legate de definirea obiectivelor și 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 acelor 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 eșantion de test, estimarea, optimizarea modelelor pentru fiecare din metodele utilizabile (validare încrucișată), compararea performanțelor.
- *Exploatare*; odată ce o metodă asociată cu un model sunt considerate ca fiind bine alese întregul eșantion este regrupat pentru a face o ultimă estimare a modelului, exploatarea modelului și difuzarea rezultatelor.

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

- Extragerea datelor, cu sau fără eșantionare, recurgând la tehnici de sondaj aplicate sau aplicabile bazelor de date.
- Explorarea datelor pentru detectarea valorilor aberante sau doar atipice, a incoerențelor, pentru studiul de distribuțiilor, structurilor de corelație, pentru căutarea tipologiilor, pentru transformarea datelor.
- Partiționarea aleatoare a eșantionului (învățare, validare, testare), în funcție de mărimea acestuia și de tehnicile care vor fi utilizate, pentru a estima o eroare de predicție în vederea alegerii modelului, a alegerii și certificării metodei.
- Pentru fiecare din metodele luate în considerație: estimarea modelului pentru o valoare dată unui parametru de *complexitate* (numărul de variabile, de vecini, de frunze, de neuroni, durata de învățare, etc.) și optimizarea acestui parametru.
- Compararea modelelor optimale obținute (câte unul pentru fiecare metodă) prin estimarea erorii de previziune.
- Iterarea eventuală a etapelor precedente, în cazul în care eșantionul de test este prea mic. Partiționări aleatoare succesive ale eșantionului pentru medierea pe mai multe cazuri a estimării finale a erorii de predicție și asigurarea robusteții modelului obținut.
- Alegerea metodei adoptate, pe baza capacităților sale de predicție, a robusteții sale dar și, eventual, a interpretabilității modelului obținut.

5. Concluzii

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

Pe de o parte, multe metode urmăresc aceleași 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 eșantionului, fie că geometric este mai bine adaptată topologiei grupurilor de discriminat, fie datorită mai buneii interacțiuni cu tipurile de variabile. Astfel, în multe situații, poate fi esențială și 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 construcția modelelor (singurele optimale).

Pe de altă parte, metodele nu prezintă toate aceleași facilități de interpretare. Nu există o cea mai bună alegere a priori. Numai experiența și un protocol de test atent construit permit determinarea acesteia. Este și motivul pentru care sistemele software generaliste nu fac o alegere și oferă aceste metode în paralel pentru a se adapta mai bine la date, la deprinderile fiecărui utilizator (client potențial) și chiar și “modei”.

În fazele exploratorii pot fi găsite relații care aparent au semnificații importante, valabile în interiorul setului de testare, dar care s-ar putea să fie fără nici o semnificație statistică într-o populație 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 totuși extrapolabile sau generalizabile la alte date decât cele studiate. Rezultatele previziunii pot fi deci viciate de o importantă eroare relativă legată de varianța estimațiilor parametrilor. Problema este de a găsi un compromis bun între bias-ul unui model mai mult sau mai puțin fals și varianța estimatorilor.

Obiectivul esențial rămâne „căutarea sensului” în vederea facilitării luărilor de decizie, prezervând fiabilitatea. Prezența sau controlul unei expertize statistice rămâne inevitabilă pentru că necunoașterea limitelor și capcanelor metodelor utilizate poate conduce la aberații de natură să discrediteze demersul, făcând caduce investițiile consimțite.

Referințe

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 și aplicații. București: Editura Academiei Române, 2009, 277 p.
4. FILIP, F. G. : Decizie asistată de calculator: decizii, decidenți – metode de bază și instrumente informatice asociate, Ed. a 2-a, rev. București: Editura Tehnică, 2005, 376 p.
5. FILIP, F. G. : Sisteme suport pentru decizii, Ed. a 2-a, rev. București: Editura 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. Cluj-Napoca: Editura Risoprint, 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.