Academia Română
Secţia Ştiinţa şi Tehnologia Informaţiei
Institutul de Cercetări pentru Inteligenţă Artificială „Mihai Drăgănescu”

mat. Niculae Cornel LEPĂDATU

CONTRIBUȚII PRIVIND ASISTAREA DECIZIILOR BAZATĂ PE DESCOPERIREA CUNOȘTINȚELOR DIN DATE

CONTRIBUTIONS TO SUPPORT THE DECISION MAKING BASED ON DATA MINING AND KNOWLEDGE DISCOVERY

- rezumatul tezei de doctorat -

Coordonator ştiinţific:

Acad. Florin Gheorghe FILIP

BUCUREȘTI - 2015

CUPRINSUL TEZEI DE DOCTORAT

INTRODUCERE

Preambul. Teza de doctorat reprezintă materializarea cercetărilor, privind domeniul asistării deciziilor, efectuate de autor în perioada 2008 – 2015 sub directa îndrumare a domnului academician Florin Gheorghe FILIP.

Conceptul nou abordat, în cadrul tezei de doctorat, este „sistemul suport pentru decizii al unei instituții de tip bibliotecă hibridă, SSD-BibHib".

Construirea SSD-BibHib este originală și se bazează pe cunoașterea, adoptarea, adaptarea și utilizarea unor instrumente științifice de actualitate: cerințele funcționale privind descrierile bibliografice (IFLA – International Federaration of Library Associations), facilitățile oferite de tehnolologia OLAP (On-Line Analitical Processing), modelarea multidimensională a bazelor de date (model evoluat, independent de orice aspecte de implementare), definirea elementelor multidimensionale (nivel minim de granularitate pentru măsuri), asigurarea sumarizabilității (definirea sistemului unitar și deschis de formule evaluabile pentru toți indicatorii de stare și de performanță ai bibliotecii), obținerea arborilor de atribute (abordare hibridă secvențială), obținerea schemei conceptuale a depozitului de date (tip „constelație"), metodele și modelele tehnologiei KDD (Knowledge Discovery from Databases) și strategia de utilizare a acestora, integrarea componentelor sistemului cu depozitul de date.

Abordarea, multidisciplinară, utilizează o arhitectură complexă de sistem suport pentru decizii, specifică, obținută prin combinarea unei tehnologii de management a bazelor de date (DBMS - Database Management System) cu două tehnologii de management a rezolvatoarelor flexibile (OLAP și KDD).

Noțiunile de bază și proprietățile lor precum și aspectele metodologice, utilizate în cadrul lucrării, sunt prezentate în capitolele de sinteză și sunt evidențiate în contextele specifice de utilizare ale acestora. Pentru fiecare din domeniile abordate sunt menționate tendințele specifice domeniului, actualitatea abordărilor și viziunile/contribuțiile personale.

Pentru comunitatea științifică, pentru cercetătorii și practicienii care abordează dezvoltarea de sisteme suport pentru decizii, modul de abordare și construire a SSD-BibHib oferă un cadru conceptual și metodologic de integrare a depozitării datelor cu analiticile on-line și data mining care se poate dovedi foarte util în demersurile lor.

Prin construcție, sistemul SSD-BibHib oferă facilități consistente de alimentator cu cunoștințe pentru diferite alte sisteme suport pentru decizii ale instituțiilor/companiilor.

Ideea, originală, de a conferi sistemului suport pentru decizii al unei biblioteci un rol de alimentator de cunoștințe pentru sistemele suport pentru decizii ale companiilor, presupune și stimulează cercetări privind realizarea de instrumente de integrare prin sinergie a diferitelor tehnologii de management a cunoștințelor astfel încât, pentru a satisface o solicitare a unui utilizator sau pentru a reacţiona la un anumit eveniment, într-o singură operaţie orice capabilitate să poată funcționa independent de oricare alta sau împreună cu oricare alta utilizând orice format de reprezentare a elementelor de cunoaștere.

Cercetările multidisciplinare efectuate se încadrează, în concordanță cu INCOSE (International Council on Systems Engineering), în domeniul științific „Ingineria sistemelor" din domeniul fundamental „Științe inginerești".

Rezultatele cercetărilor efectuate pe parcursul elaborării tezei de doctorat se regăsesc în referatele doctorale (vizibile pe Internet) și sunt deja publicate în reviste de specialitate.

Prolog. Pe măsura dezvoltării societăţii omeneşti, dezvoltarea managementului s-a impus ca un proces de orientare a activităţilor umane în vederea atingerii obiectivelor dorite. Multă vreme managementul a fost considerat ca fiind o artă, el având la bază intuiţia, raționamentul, creativitatea, experienţa şi cunoştinţe dobândite, mai mult prin încercări sau erori decât prin metode cantitative susţinute de o abordare ştiinţifică.

Mediul economic, social şi politic în care se iau în prezent deciziile manageriale se caracterizează printr-o dinamică pronunţată şi continuă în care tehnologiile avansate devin un determinant major al stilului de viaţă uman. Pentru managerii actuali numărul căilor de acţiune posibile poate fi foarte mare, gradul de incertitudine poate face foarte dificilă previziunea consecinţelor luării unei decizii, efectele unor erori în luarea deciziilor ar putea fi dezastruoase datorită complexităţii operaţiilor şi reacţiilor în lanţ pe care aceste erori pot să le cauzeze.

Convergenţa procesării informaţiei cu tehnicile de comunicaţii, ilustrată elocvent mai ales prin dezvoltarea exponenţială a Internet-ului, a determinat apariţia unor enorme cantităţi de date, informaţii şi cunoştinţe reprezentate în forme din cele mai diverse. Această cantitate imensă de informații este sporită, în continuu, nu doar de dezvoltările permanente ale *web*-ului dar şi de apariţia agresivă a unor tehnologii emergente precum sistemele dedicate (*embeded*), sistemele mobile şi respectiv sistemele omniprezente (*ubiquitous*) de prelucrare a informaţiei.

Este, deci, indiscutabil de clară necesitatea extragerii de informaţii şi de cunoştinţe, din aceste masive de date distribuite, în primul rând pentru asistarea proceselor decizionale. In acest sens, esenţial este faptul că este nevoie de a reprezenta în mod explicit caracteristici importante ale informaţiilor, care nu mai sunt legate de reprezentarea abstractă a conceptelor lumii reale ci, mai degrabă, de obiectivul factorilor de decizie şi anume susţinerea proceselor de analiză a datelor orientate către luarea deciziilor.

O organizație poate dispune de sisteme/subsisteme informatice specifice funcțiilor sale precum și nivelurilor managementului său, strategic, tactic și operațional. Unele dintre aceste sisteme informatice, întâlnite în literatura de specialitate sub denumirea de sisteme suport pentru decizii, s-au impus în deservirea nivelelor strategic și tactic și evoluția pe termen mediu și lung a organizației.

Menirea sistemelor suport pentru decizii este de a atenua efectul limitelor şi resticţiilor factorului decizional în rezolvarea problemelor decizionale. În desfăşurarea proceselor decizionale poziţia centrală este ocupată de intuiţia şi judecata umană iar metodele utilizate se bazează pe analiza datelor disponibile.

Principalele concepte şi rezultate în domeniul asistării cu mijloace informatice a activităţilor din procesele decizionale, care presupun analiza datelor, au provenit din prelucrarea analitică on-line (*on-line analytical processing*) și depozitarea datelor (*data warehousing*) precum şi din explorarea datelor şi descoperirea cunoştinţelor (*data mining and knowledge discovery*).

Arhitectura generică a sistemelor suport pentru decizii se compune din patru componente esenţiale: un sistem de limbaj, format din mesaje pe care sistemul le poate accepta; un sistem

de prezentare, format din mesaje pe care sistemul le poate emite; un sistem al elementelor de cunoaştere, constând din cunoştinţe deţinute de sistem şi, în fine, un sistem de tratare a problemei, constând din module software prin care elementele de cunoaştere sunt procesate ca urmare a interpretării mesajelor de intrare.

Arhitecturile personalizate păstrează caracteristicile sugerate de arhitectura generică dar sunt orientate către o anumită tehnologie de reprezentare şi prelucrare a cunoştinţelor. Dacă factorul decizional are nevoie de capacităţile de prelucrare oferite de mai multe tehnologii de management a cunoştinţelor acesta poate opta, fie pentru utilizarea mai multor sisteme suport pentru decizii, fiecare orientat către o anumită tehnologie, fie pentru utilizarea unui singur sistem suport pentru decizii, dar care integrează mai multe tehnologii.

Arhitectura generică a sistemelor suport pentru decizii evidențiază modul în care cele patru componente sunt legate atât între ele cât şi cu utilizatorul sistemului. Utilizatorul este, de obicei, un factor de decizie sau un participant la luarea deciziei dar, deasemenea, poate fi atât un dezvoltator sau un administrator al sistemului cât şi un alimentator de cunoștințe, persoană sau dispozitiv care furnizează sistemului suport pentru decizii date, informații sau cunoștințe de intrare.

Sistemele informatice ale bibliotecilor (*integrated library system - ILS*) pot și tind să devină, în mod natural, actori foarte importanți în alimentarea cu cunoștințe a sistemelor suport pentru decizii ale organizațiilor. Sprijinul bibliotecilor și bibliotecarilor în luarea deciziilor a variat în timp, de la unul pasiv (colecţii tradiţionale de cărţi şi reviste ale bibliotecii) către unele extrem de active (asistenți decizionali).

Bibliotecile digitale au oferit perspective noi pentru sistemele suport pentru decizii ale organizațiilor. Având în vedere imensitatea volumului de informații care se acumulează în bibliotecile digitale, unul dintre cei mai imperativi parametri de implementare a unui scenariu de extragere/generare orientată către cerințe a cunoștințelor este *data mining*. *Bibliomining* (termen derivat din *bibliometrics* și *data mining*) a oferit perspectiva ca, prin intermediul unui singur depozit de date, să se prelucreze cunoștințe privind interconexiunile dintre rețele sociale diferite, respectiv, dintre comunitatea formată de autori și comunitatea formată de bibliotecă și utilizatorii săi.

Demersul de realizare al unui sistem suport pentru decizii de bibliotecă cu rol de alimentator de cunoștințe, nou și captivant, creează multe provocări dar promite mari îmbunătățiri în modul de desfășurare a activităților, în modul de înțelegere a ceea ce se face în prezent și a ceea ce se preconizează pentru viitor.

Obiective. Prezenta teză și-a propus drept obiective :

- Studiul funcționalităților și modalităților de utilizare a tehnologiei *data mining* în procesele decizionale.
- Construirea unui sistem suport pentru decizii în domeniul bibliotecilor menit:
  - să susțină procesele decizionale din biblioteci hibride, inclusiv prin utilizarea optimală a capabilităților oferite de tehnologia *data mining*;
  - să îndeplinească un rol consistent de alimentator cu cunoștințe, atât pentru sine cât și pentru sistemele suport de decizii ale altor organizații;
  - să ofere un model de abordare în construirea sistemelor suport pentru decizii.

Organizare. Teza este organizată după cum urmează:

- Capitolul 1 - Sistemele suport pentru decizii
- Capitolul 2 - Tehnologia *data mining*
- Capitolul 3 - Alimentarea cu cunoștințe a sistemelor suport pentru decizii

Capitolul 1 reprezintă o sinteză, generală, referitoare la abordarea sistemică și multidisciplinară a situaţiilor decizionale, cu focalizare pe deciziile manageriale și adoptarea deciziilor prin metode ştiinţifice, dintro perspectivă modernă și cu un grad ridicat de conceptualizare şi de generalitate. Aspectele abordate privesc: mediul decizional din zilele noastre, informatica decizională, arhitecturi și tehnologii care susțin realizarea și utilizarea sistemelor suport pentru decizii. Referitor la mediul decizional au fost evidențiate: viziunea sistemică privind managementul organizațiilor, situaţiile şi problemele decizionale, tipologia deciziilor şi rolurile decizionale, procesele decizionale, asistenţii decizionali și asistarea deciziilor în mediile informatizate. Referitor la informatica decizională, termen distinct de informatica de gestiune, au fost evidențiate: eforturile de definire a conceptului de sistem suport pentru decizii, caracteristicile și funcţiunile unui astfel de sistem informatic, taxonomia, unele evaluări strict necesare, dinamica preocupărilor teoretice și aplicative privind această clasă de sisteme. Referitor la arhitecturile sistemelor suport pentru decizii au fost evidențiate: arhitectura generică, arhitecturi personalizate și arhitecturi integrate precum și arhitectura preferată în contextul tezei. În ceeace privește tehnologiile s-au evidențiat unele aspecte esențiale, conceptuale și metodologice, referitoare la: procesul de construire al unui sistem suport pentru decizii, modelarea multidimensională a datelor și proiectarea conceptuală a depozitelor de date, explorarea datelor și descoperirea cunoştinţelor (problemele rezolvabile, produsele software suport, strategia de aplicare și efervescența utilizărilor) în concordață cu arhitectura aleasă. Au fost, de asemenea, evidențiate elementele de noutate și implicațiile științifice, tehnologice, economice și sociale pentru fiecare din aspectele abordate.

Capitolul 2 reprezintă o sinteză, din perspectiva prospectorului de date, a variantelor recente de soluții informatice realizate pentru metodele/modelele cele mai frecvent utilizate în aplicațiile *data mining*. Structurarea materialului s-a bazat pe cele două demersuri, exploratoriu și explicativ, precum și pe strategia de aplicare *data mining*. Referitor la soluțiile exploratorii dintre metodele de analiză factorială au fost evidențiate: analiza în componente principale, analiza factorială discriminantă, analiza corespondenţelor simple, analiza corespondenţelor multiple și analiza canonică. Dintre metodele de analiză a grupurilor au fost evidențiate: abordarea ierarhică, abordarea neierarhică (partițională) și abordarea mixtă. Referitor la soluțiile exploratorii au fost evidențiate: modelele liniare (de analiză a regresiei, de analiză dispersională și generalizate), metodele de discriminare (geometrice și probabiliste), mașinile cu suport vectorial, metodele conexioniste (rețelele neuronale), metodele de segmentare (arborii de clasificare și regresie) și metodele de agregare a modelelor. De asemenea, pentru fiecare metodă/model au fost evidențiate, după caz, o serie de aspecte specifice esențiale, necesare prospectorului de date, precum: spațiile de reprezentare, semnificaţii ale coeficienţilor, puterea de discriminare a caracteristicilor, metode de selecţie a variabilelor, domenii de aplicabilitate, gradul estimat de adecvare la datele observate, măsurarea performanțelor, separarea estimării modelului de estimarea erorilor de previziune, controlul supraajustării, elemente de noutate și performanțe computaționale, caracterizarea și interpretarea rezultatelor, relaţii care pot exista cu alte metode/modele pentru situații în care devine oportună o utilizare combinată.

Capitolul 3 reprezintă contribuția propriu-zisă a tezei la evoluțiile din domeniul sistemelor suport pentru decizii. Ideea abordării este, în primul rând, de a defini un sistem informatic, menit să susțină procesele decizionale aferente unei anumite categorii de organizații, capabil să integreze, cât mai fiabil, aplicațiile necesare de explorare a datelor și descoperire a cunoștințelor și pe care să le exploateze cât mai eficient posibil. În al doilea rând, orientarea către organizațiile de tip bibliotecă este de natură să permită ca sistemul să fie conceput astfel încât să poată îndeplini, inclusiv, un rol consistent de alimentator de cunoștințe atât pentru sine cât și pentru diverse sisteme suport pentru decizii ale altor companii. Cercetările întreprinse au vizat: identificarea modurilor de implicare a bibliotecilor în susținerea activităților decizionale din diverse organizații, descrierea caracteriticilor și funcționalităților sistemului suport pentru decizii al bibliotecii, analiza cerințelor informaționale, proiectarea concepuală a depozitului de date, elaborarea și experimentarea de proceduri pentru descoperirea/generarea de cunoștințe. Referitor la implicarea bibliotecilor în sprijinirea activităților decizionale ale altor organizații au fost evidențiate participările bibliotecarilor în calitate de asistenți decizionali și perspectivele oferite de bibliotecile digitale *text-mining, web-mining, bibliomining.* Referitor la sistemul suport pentru decizii al bibliotecii au fost evidențiate și descrise: obiectivele și arhitectura sistemului, direcțiile posibile de îmbunătățire a activităților și avantajele oferite de sistem, modalități de realizare, etape și resurse necesare. Referitor la etapa de analiză a cerințelor informaționale au fost evidențiate, minuțios analizate și formalizate, cerințele bibliografice, biblioteconomice, bibliometrice și de reconciliere cu sursele de date. Noțiunile și conceptele introduse au permis obținerea de definiții evaluabile pentru toți indicatorii uzuali, în strictă concordanță semnificațiile curente ale acestora, rezultând un singur sistem de indicatori, unitar, integrat și deschis. Referitor la proiectarea conceptuală a depozitului de date au fost identificate și descrise: subiectele majore de interes ale factorilor de decizie (faptele); perspectivele de analiză pentru fiecare din faptele identificate (dimensiunile); aspectele specifice şi măsurabile ale faptelor, relevante pentru analiză (măsurile). Au fost elaborate, de asemenea, modelul multidimensional al datelor (arborii de atribute sau cuburile de date) și schema conceptuală (tip constelație) a depozitului de date. Referitor la descoperirea/generarea de cunoștințe noi prin analiza datelor stocabile în depozitul de date au fost avute în vedere proceduri de importanță majoră pentru factorii decizionali ai unei biblioteci fiind evidențiate: identificarea și ierarhizarea preferințelor de lectură ale utilizatorilor bibliotecii, identificarea și ierarhizarea subiectelor de interes pentru utilizatori, identificarea și ierarhizarea autorilor pe diferite subiecte de interes, gruparea documentelor în funcție de conținut, recomandarea documentelor recent achiziționate în funcție de profilurile utilizatorilor. Datele, necesare acestor proceduri, s-au regăsit în depozitul de date, fiind obtenabile din sursele de date, fapt datorat analizei preliminare, riguroase și complete, a cerințelor informaționale ale sistemului. Algoritmii de prelucrare au putut fi mult simplificați, datorită modelării multidimensionale avansate a datelor, iar performanțele proceselor computaționale sunt susținute de prelucrările analitice on-line.

Mulțumiri. Autorul mulțumește, în primul rând, domnului academician Florin Gheorghe FILIP, conducătorul științific al tezei, pentru oportunitățile oferite, îndrumările și sprijinul acordat în atingerea obiectivelor urmărite, pe tot parcursul perioadei de doctorat .

Autorul mulțumește, de asemenea, tuturor colegilor cu care a colaborat în această perioadă prin schimb de idei, de cunoștințe sau soluții cât și prin realizarea proiectului complex „Sisteme suport pentru cultura cunoașterii bazate pe soluții și instrumente din domeniul

Business Inteligence – SSCBI" (Contract CEX-05-D8-19/2005, derulat în perioada 2005 – 2008, http://sscbi.ici.ro/Contact.htm): dr. Măriuca Stanciu și fil. Gabriela Dumitrescu de la Biblioteca Academiei Române; dr. Cristina Niculescu și dr. Angela Ioniță de la Institutul de Cercetări pentru Inteligenţă Artificială „Mihai Drăgănescu" al Academiei Române; dr. Ioan Stancu-Minasian, dr. Voicu Boșcaiu, dr. Cornelia Enăchescu, dr. Denis Enăchescu și dr. Viorel Vodă de la Institutul de Statistică Matematică și Matematică Aplicată "Gheorghe Mihoc - Caius Iacob" al Academiei Române; mat. Cornelia Ioana Lepădatu, mat. Dora Coardoș, dr. Vasile Coardoș† și ing. Alexandru Marinescu de la Institutul Național de Cercetare-Dezvoltare în Informatică, ICI București.

Autorul mulțumește, totodată, doamnei Cornelia Ioana Lepădatu, în calitate de soție, pentru răbdare, empatie, înțelegere și sprijin.

Capitolul 1. SISTEME SUPORT PENTRU DECIZII

Mediul decizional. Managementul organizațiilor, definit ca fiind aplicarea metodei ştiinţifice în analiza şi soluţionarea problemelor de decizie managerială, se caracterizează prin abordarea sistemică și multidiciplinară a situaţiilor decizionale, focalizarea pe deciziile manageriale și adoptarea deciziilor prin metode ştiinţifice, folosirea modelelor matematice formale și utilizarea pe scară largă a tehnologiilor informaţiei şi comunicaţiilor. Sunt definite conceptul teoretic de sistem, problema abstractă de management și obiectivele managementului, situațiile și problemele decizionale, conținutul și tipologia deciziilor, actorii implicați în luarea deciziilor precum și rolurile decizionale ale acestora.

Pentru procesul decizional, constituit dintr-o succesiune de activităţi decizionale, sunt prezentate: modelul procesual, modelul cel mai larg acceptat pentru reprezentarea desfăşurării activităţilor decizionale, cu cele patru faze ale sale: informarea (*intelligence*), proiectarea variantelor şi modelelor (*design*), alegerea (*choice*), implementarea şi evaluarea rezultatelor (*review*) precum și modelul bazat pe cunoaștere, consistent cu cel procesual, menit să exprime, din perspectiva modernă a prelucrării cunoştinţelor și cu un grad ridicat de conceptualizare şi generalitate, modul de desfăşurare a activităţilor decizionale.

În activităţile decizionale, decidenţii sunt ajutați cu anumite entităţi de suport precum asistenţii decizionali şi instrumentele informatice. Conceptul de sistem uman de suport pentru decizii (*Human Decision Support System*) este menit să descrie activitatea asistenţilor decizionali utilizând elementele de cunoaştere (baza de cunoștințe) care se referă atât la domeniul aplicaţiei şi universul decidentului asistat cât şi la instrumentele, procedurile şi raţionamentele care sunt necesare pentru rezolvarea problemelor decizionale. Insrumentele informatice au parcurs o serie de etape de dezvoltare istorică, cele mai semnificative fiind: sistemele de prelucrare automată a datelor (*Automatic Data Processing*), sistemele tranzacţionale (*Transaction Processing Systems*), sistemele de informare a conducerii (*Management Information Systems*), sistemele suport pentru decizii (*Decision Support Systems*). Aceste instrumente coexistă și conlucrează între ele, altele (de analiză, modelare, optimizare, simulare sau inteligență artficială) susțin în mod specific numai anumite activități din procesul decizional.

Informatica decizională. Sistemele suport pentru decizii formează o clasă eterogenă de sisteme informatice antropocentrice, adaptive şi evolutive, care interacţionează cu celelalte părţi ale sistemului informatic al organizaţiei și au menirea de a atenua efectul limitelor decidentului intelectual. O taxonomie a acestor sisteme este necesară şi utilă în multiple

scopuri, principalele clasificări se bazează în esență pe tipologiile deciziilor și decidenților și utilizează drept principale criterii tipul decidentului, tipul de suport și orientarea tehnologică.

Sunt evidențiate sintetic situațiile în care considerarea introducerii unui sistem suport pentru decizii este oportună adică situațiile în care investiţia se poate dovedi justificată şi profitabilă, principalele condiţii necesare care dacă sunt îndeplinite se constituie în premise ale succesului, principalele efecte benefice obtenabile, principalele limite, principalele cauze posibile de insucces și bineînțeles necesitatea unei bune motivări a deciziei de introducere a sistemului suport pentru decizii.

Aplicaţiile practice au confirmat multe din speranţele care au însoţit apariţia conceptului de sistem suport pentru decizii. S-a constatat că succesul sistemelor suport pentru decizii a fost determinat nu numai de calităţile tehnice ale sistemului, ci şi de „buna potrivire" a acestuia atât cu aptitudinile şi cunoştinţele utilizatorului cât şi cu caracteristicile situaţiilor decizionale.

Utilizarea sistemelor suport pentru decizii s-a răspândit în toate domeniile de activitate, ele au evoluat în timp sub influența evoluțiilor tehnologice și organizaționale. Interesul oamenilor de știință pentru sistemele suport pentru decizii a crescut de-a lungul anilor, evoluția din ultimele decenii a materialelor publicate denotă o creștere aproape exponențială a preocupărilor privind această clasă de sisteme informatice.

***Arhitectura sistemelor suport pentru decizii*.** Între sistemele suport pentru decizii există diferenţieri semnificative determinate de domeniile de aplicabilitate, de caracteristicile de utilizare, de funcţionalităţile proiectate, de abordările privind interacţiunile dintre componente, de modalităţile de încorporare în procesele decizionale, de tipurile de beneficii rezultate din utilizare.

Conturarea unui cadru conceptual capabil să acopere majoritatea soluţiilor arhitecturale identificabile în sistemele suport de decizie specifice, a fost favorizată de evoluția ideilor privind conceptele de sistem uman suport pentru decizii și de procesor pentru probleme decizionale, privind modelul bazat pe cunoaştere al activităţilor decizionale și de ideile privind extensiile sistemelor de gestiune a bazelor de date pentru a integra date cu modele.

Prin prisma arhitecturii generice, orice sistem suport pentru decizii se compune din patru componente esenţiale (subsisteme): de limbaj, reprezentând mulţimea formelor de exprimare prin care utilizatorul poate transmite solicitări (mesaje de intrare) ce pot fi înţelese şi acceptate de către sistem, sau prin care terţi îşi transmit rapoarte; de prezentare, reprezentând mulţimea formelor şi mijloacelor prin care sistemul emite mesaje de ieşire către utilizator sau către terţi; de cunoştinţe, conţinând elementele de cunoaştere achiziţionate sau create în interiorul sistemului; de tratare a problemei decizionale, reprezentând mulţimea modulelor software prin care elementele de cunoaştere disponibile sunt prelucrate ca urmare a interpretării mesajelor de intrare. Utilizatorul sistemului suport pentru decizii este de obicei un factor de decizie sau un participant la luarea deciziei dar, deasemenea, poate fi atât un dezvoltator sau un administrator al sistemului cât şi un alimentator de cunoștințe – persoană sau dispozitiv care furnizează date / informații / cunoștințe de intrare.

Arhitecturile personalizate păstrează caracteristicile sugerate de cadrul generic dar sunt specializate pe o anumită tehnologie (sau tehnologii) de reprezentare şi prelucrare de cunoştinţe (texte, hypertext, baze de date, foi electronice de calcul, rezolvatoare, reguli, tehnologii combinate). Dacă factorul decizional are nevoie de capacităţile de prelucrare oferite de mai multe tehnologii de management al cunoştinţelor poate opta fie pentru

utilizarea mai multor sisteme suport pentru decizii, fiecare orientat către o anumită tehnologie fie pentru utilizarea unui singur sistem suport pentru decizii, dar care integrează mai multe tehnologii.

Arhitectura sistemelor suport pentru decizii descrisă și utilizată în prezenta lucrare reprezintă un caz special de integrare, deosebit de important prin implicaţiile sale și a rezultat din combinaţia dintre o tehnologie de management a bazelor de date şi o tehnologie de management a rezolvatoarelor flexibile. Combinarea depozitării datelor cu rezolvatoarele analitice este foarte utilizată în prezent de către companii pentru a obţine noi informaţii în timp ce combinarea depozitării datelor cu rezolvatoarele *data mining* poate genera cunoştinţe noi, deosebit de utile în luarea deciziilor, prin descoperirea de *pattern*-uri din date.

Tehnologia sistemelor suport pentru decizii. Procesul de construire al unui sistem suport pentru decizii specific de aplicaţie se compune din o serie de activităţi care încep cu generarea ideii de introducere a sistemului în organizaţie şi se termină cu obţinerea unei versiuni relativ stabile, utilizabile în mod curent, a sistemului. Metodologic activităţile care compun procesul de construire sunt grupate în etape (cu rezultate specifice), care corespund ciclului de viaţă al oricărui sistem informatic, iniţierea şi pregătirea proiectului (studiul de fezabilitate); analiza de sistem (specificaţia de detaliu); proiectarea tehnică (proiectul de execuţie); implementarea (sistemul operaţional); exploatarea (luări de decizii); evoluţia (perfecţionarea sistemului).

Prelucrarea analitică on-line permite analiștilor și managerilor să înleleagă esența datelor prin acces rapid, consistent şi interactiv la o mare varietate de vederi posibile reflectând dimensiunile reale ale unei organizații. Modelarea conceptuală multidimensională a datelor intervine în etapele inițiale ale procesului de construire al unui sistem suport pentru decizii pentru a defini cerinţele în cel mai bun mod posibil. Modelul de date multidimensional ales și (re)definit riguros satisface cerinţele fundamentale pe care orice model multidimensional trebuie să le îndeplinească în contextul aplicaţiilor *OLAP* precum și o serie de caracteristici suplimentare, recomandate şi considerate avansate. Reprezentările conceptuale multidimensionale, care nu mai sunt legate de reprezentarea abstractă a conceptelor lumii reale ci de susţinerea proceselor de analiză a datelor orientate către luarea deciziilor, furnizează o descriere în termeni abstracţi a conţinutului depozitului de date utilizată ca referinţă în conceperea interogărilor analitice complexe.

Proiectarea conceptuală a depozitului de date este pasul cel mai important în reprezentarea corectă a unui domeniu de interes, fiind elementul esențial asupra căruia atât factorii de decizie cât şi informaticienii sunt de acord. Pentru proiectanți este foarte important să urmeze o abordare specifică, consolidată şi robustă dat fiind că dezvoltarea unui depozit de date este un proces foarte costisitor chiar şi astăzi când există multe instrumente software acoperind toate etapele din ciclul de viaţă al depozitului de date oferind chiar și soluții prefabricate. Proiectarea conceptuală a unui depozit de date poate fi obținută prin mai multe categorii de metode: orientate către date (*data-driven*), metode orientate către cerinţe (*goal-oriented*) și metode mixte (*hybrid*). Deoarece primele două categorii de metode sunt în antiteză una cu cealaltă, proiectanții fiind nevoiți să aleagă una dintre ele, este preferabilă o metodă din a treia categorie care, eventual, să remedieze din neajunsurile și să valorifice din avantajele fiecăreia. Metoda aleasă, dezvoltată și urmată în lucrare, combină şi integrează (*integration-derived*) o fază de abordare orientată către cerinţe cu o fază de abordare orientată către date, cele două faze sunt executate secvențial ieşirea primei faze fiind utilizată ca intrare în a doua fază. În esență, etapele generale ale metodei sunt: analiza cerinţelor; modelarea multidimensională a

datelor; reconcilierea cu sursele de date; definirea arborilor de atribute (cuburilor de date); modelarea avansată a datelor.

Explorarea datelor şi descoperirea cunoştinţelor (mineritul datelor sau *data mining*) oferă un ansamblu de metode şi algoritmi destinat explorării şi analizei unor (adesea) mari volume de date în vederea deducerii, din aceste date, a unor asocieri, a unor reguli, a unor tendinţe necunoscute (nefixate a priori), a unor structuri specifice care să restituie în mod concis esenţa informaţiei utile pentru asistarea deciziilor. Conceptele, metodele şi tehnicile oferite de *data mining* sunt relativ vechi, dezvoltarea acestora în decursul timpului se încadrează în trei perioade istorice distincte (statistică, analiza datelor, explorarea datelor și descoperirea cunoștințelor) fiecare perioadă fiind definită prin aspectele caracteristice ale utilizării. *Data mining* nu este, deci, nici noutate tehnologică nici ştiinţifică, noutatea a constat în integrarea acestei tehnologii în procesarea industrială a informaţiei. Aportul *data mining* se rezumă la un număr limitat de acţiuni care, folosite în mod adecvat, se dovedesc extrem de utile pentru numeroase probleme şi situaţii din domeniul decizional. Pentru principalele tipuri de probleme rezolvabile cu *data mining,* cele mai frecvente fiind analiza asocierilor, *pattern*-urile secvenţiale, analiza grupurilor, clasificarea, mulţimile *rough* și *link mining,* sunt prezentate unele variante de definire formalizată ale acestora.

Utilizarea *data mining* presupune: evaluarea oportunităţii acesteia şi identificarea datelor pe care se poate baza explorarea; extragerea de informaţii din colecţiile/depozitele de date existente şi prelucrarea acestora prin metode/tehnici adecvate de *data mining*; adoptarea de decizii pe baza rezultatelor obţinute şi întreprinderea de acţiuni; măsurarea rezultatelor concrete pentru a identifica şi alte modalităţi de exploatare a datelor. Ceea ce se exploatează prin *data mining* sunt colecţii de date disponibile, provenite din surse interne ale organizaţiei care au fost constituite, inclusiv ca structură, în perspectiva altor finalităţi la care se adaugă date provenite din diverse alte surse externe organizaţiei. Datele utilizate sunt privite ca reprezentând o serie de observaţii privind o mulțime de caracteristici (sau variabile) care au fost măsurate pe o mulțime (populație sau eșantion) de obiecte (sau indivizi).

Utilizarea *data mining* este susținută de numeroase platforme de produse informatice dedicate, unele realizate pentru *PC*-uri (simplu de instalat, conviviale, cu algoritmi de bună calitate şi nu foarte scumpe) menite să exploateze volume suficient de mari de date şi oferind în general una sau două tehnici altele, putând funcţiona pe arhitecturi de tip client-server, menite să exploateze volume foarte mari de date, cu o paletă largă de tehnici atât în variantă statistică cât şi în variantă *data mining*. Sunt prezentate principalele tipuri de cerinţe privind un software de *data mining,* cele mai utilizate produse informatice în funcţie de volumele de date prelucrate precum și o evaluare relevantă (prin distribuțiile și evoluțiile anuale, din perioada 2008 – 2012, pe 29 domenii de activitate) a efervescenței utilizărilor tehnologiei *data mining* rezultată din *KDnuggets Polls on Data Mining Application*.

Utilizarea *data mining* trebuie făcută conform unei strategii de utilizare, simple, clare și pragmatice care depinde în esenţă de tipurile de variabile considerate (explicative și/sau de explicat) de natura variabilelor (cantitative și/sau calitative) şi de obiectivele urmărite (explorare multidimensională sau reducere de dimensiune, clasificare sau segmentare, modelare sau discriminare) și care constă în înlănţuirea a patru etape majore: extracţie, explorare, analiză și exploatare. Fundamentul utilizării tehnologiei *data mining* îl constituie succesiunea a două demersuri: primul, descriptiv și exploratoriu, care se sprijină pe noţiuni elementare (medie şi dispersie), pe tehnici descriptive multidimensionale şi pe reprezentări grafice și cel de al doilea, inferenţial şi confirmatoriu, care face apel la metode destinate să

explice apoi să prevadă, urmând anumite reguli de decizie, o variabilă privilegiată cu ajutorul uneia sau mai multor variabile explicative permițând validarea (sau infirmarea) ipotezelor formulate à priori (adică urmare a unui demers exploratoriu) şi extrapolarea acestora de la nivelul unui eşantion la cel al unei populaţii mai largi.

Capitolul 2. TEHNOLOGIA DATA MINING

Soluții informatice exploratorii

Explorarea datelor se bazează pe un set de metode descriptive, în cea mai mare parte geometrice, al căror instrument matematic major este algebra matricială şi care se exprimă fără să presupună à priori un model probabilist, este destinată descrierii şi analizei datelor multidimensionale şi poate fi utilizată în orice domeniu atunci când datele sunt mult prea multe pentru a mai putea fi înţelese de o minte omenească. Aceste metode permit, în special, prelucrarea şi sinteza informației din tabelele de date de mari dimensiuni pe baza estimării corelaţiilor dintre variabilele studiate iar instrumentele statistice utilizate sunt matricea corelaţiilor și/sau matricea de varianță-covarianță. Unele dintre metode ajută la evidențierea relaţiilor care pot exista între diferite date şi elaborează informaţii statistice care permit o descriere mai succintă a informaţiei conţinute în aceste date, altele permit regrupări ale datelor în scopul de a face să apară clar ceea ce le face omogene şi astfel de a le înţelege şi de a le defini mai bine.

În demersul descriptiv şi exploratoriu obiectivele principale urmărite sunt: analiza factorială sau explorarea multidimensională, bazată cel mai frecvent pe metode precum analiza în componente principale (*ACP*), analiza factorială discriminantă (*AFD*), analiza corespondenţelor simple (*ACS*), analiza corespondenţelor multiple (*ACM*) sau analiza canonică (*AC*) și analiza grupurilor sau clasificarea, utilizând cel mai adesea metode precum clasificarea ascendentă ierarhică (*CAI*), metoda norilor dinamici (*MND*) sau metoda de clasificare mixtă (*MCM*).

Analiza factorială sau explorarea multidimensională

Spațiul variabilelor, spațiul indivizilor, proximități. Mulţimii de observaţii disponibile i se asociază o matrice $\boldsymbol{X} = \{x_{ij}\}_{i=1}^{n}{}_{j=1}^{p} \in \mathcal{M}_{n\times p}(\mathcal{R})$, unde n reprezintă numărul de indivizi, p reprezintă numărul de variabile iar elementul x_{ij} reprezintă valoarea variabilei j măsurată pe individul i. Vectorii-coloană ai matricii $\boldsymbol{X}$ definesc un nor de p puncte-variabile în $\mathcal{R}^n$ iar vectorii-linie definesc un nor de n puncte-indivizi în $\mathcal{R}^p$. Fiecare punct-individ, definit de p coordonate corespunzând valorilor celor p variabile măsurate pe acest individ, aparține unui spaţiu vectorial $\mathcal{E} \subset \mathcal{R}^p$ numit spaţiul indivizilor, de asemenea, fiecare punct-variabilă, definit de n coordonate corespunzând celor n valori ale variabilei j măsurată pe cei n indivizi, aparține unui spaţiu vectorial $\mathcal{F} \subset \mathcal{R}^n$ numit spaţiul variabilelor. Ipoteza fundamentală a unui demers exploratoriu este aceea că întreaga informaţie este conţinută în distanţele dintre punctele unui nor, respectiv dispersia punctelor din nor. În timp ce în spaţiul indivizilor interesează distanţele dintre puncte, în spaţiul variabilelor interesează unghiurile dintre ele deoarece proximitățile dintre punctele-variabile se interpretează în termeni de corelaţii.

În *analiza în componente principale* obiectivele urmărite sunt: reducerea dimensiunii (sau compresia), respectiv aproximarea matricii $\boldsymbol{X}$ de rang p printr-o matrice de rang $q \ll p$; reprezentarea grafică „optimală” a indivizilor, minimizând deformările norului de puncte, într-un subspaţiu $\mathcal{E}_q$ de dimensiune $q \leq 3$ pentru a face posibilă vizualizarea, precum și

reprezentarea grafică a variabilelor într-un subspaţiu $\mathscr{F}_q$ explicitând „cel mai bine" legăturile iniţiale între aceste variabile. În funcţie de provenienţă, variabilele care pot face obiectul unei analize în componente principale pot lua valori cantitative obţinute în urma unor măsurători, pot lua valori calitative obţinute în urma unor notaţii dar care sunt asimilabile cu variabilele cantitative sau pot lua valori calitative ordinale obţinute în urma unor clasamente dar care pot fi transformate în variabile continue. Pentru prospectorul de date aspectele cele mai interesante sunt: reprezentarea indivizilor, reprezentarea variabilelor, relaţiile de tranziţie între spaţii, reconstituirea datelor iniţiale, reprezentarea simultană a indivizilor și variabilelor și în special, interpretarea rezultatelor. Analiza în componente principale are un rol esenţial fiind metoda care servește drept fundament teoretic şi pentru celelalte metode de explorare multidimensională numite factoriale.

În *analiza factorială discriminantă* o variabilă calitativă T cu q modalităţi, jucând rolul de variabilă de explicat, generează o partiţie a celor n indivizi în q clase. În anumite situaţii se poate constata că puterea de discriminare a caracteristicilor (axelor) este slabă pentru datele considerate, fie că nu s-au ales cele mai bune caracteristici ale datelor, fie că datele sunt prin natura lor foarte asemănătoare. Pentru astfel de situaţii este uneori posibilă determinarea unui nou sistem de coordonate faţă de care structura de clase este mai evidentă decât în sistemul iniţial, axele noului sistem având o putere de discriminare a claselor superioară celei a axelor iniţiale. Determinarea axelor cu puterea de discriminare cea mai bună permite apoi definirea funcțiilor de discriminare respectiv a suprafețelor și regiunilor de decizie. Într-o analiză factorială discriminantă se disting, în consecinţă, două demersuri: primul, descriptiv, ce constă în căutarea funcţiilor de discriminare liniare pe eşantionul de volum n respectiv găsirea combinaţiilor liniare de variabile explicative ale căror valori separă cel mai bine cele q clase; al doilea, decizional, ce constă în aflarea claselor de afectare a n' indivizi noi, descrişi prin variabilele explicative.

În *analiza corespondențelor simple* două variabile calitative, T' și T'', cu τ' și, respectiv, τ'' modalități, observate simultan pe același eșantion de n indivizi, generează fiecare câte o partiţie a eşantionului. Tabelul de contingenţă, asociat analizei corespondențelor simple, este o matrice $\boldsymbol{K} = \{ k_{m'm''} \}^{\tau'}_{m'=1}{}^{\tau''}_{m''=1} \in \mathcal{M}_{\tau' \times \tau''}(\mathscr{R})$ cu τ' linii, τ'' coloane şi elementele $k_{m'm''}$, unde elementul $k_{m'm''}$ reprezintă numărul de indivizi din eșantion având simultan modalitatea m' a variabilei T' şi modalitatea m'' a variabilei T''. Analiza corespondenţelor simple se aplică tabelelor de contingenţă şi tratează în mod echivalent atât liniile cât şi coloanele. Abordările cele mai recente constau în a defini analiza corespondenţelor simple ca fiind rezultatul a două analize în componente principale, pentru profiluri-linii şi pentru profiluri-coloane, utilizând metrica χ^2.

În *analiza corespondențelor multiple* se dispune de observaţii privind $s > 2$ variabile calitative $\{T^q\}^s_{q=1}$, fiecare variabilă având, respectiv, câte $\{\tau_q\}^s_{q=1}$ modalităţi; variabilele au fost observate simultan pe un eşantion de n indivizi, fiecare individ alegând una şi numai una dintre modalităţile fiecărei variabile; modalităţile fiecărei variabile se exclud reciproc, fiecare modalitate fiind observată cel puţin o dată.

Notând cu $\tau = \sum^s_{q=1} \tau_q$ numărul total de modalităţi ale celor s variabile şi cu $m_{iq} \leq \tau_q$ numărul modalităţii alese de individul i, dintre cele τ_q modalităţi ale variabilei T^q, tabelul de date condensat, respectiv matricea $\boldsymbol{M} = \{m_{iq}\}^n_{i=1}{}^s_{q=1} \in \mathcal{M}_{n \times s}(\mathscr{R})$, descrie cele s modalităţi alese de cei n indivizi dar nu este exploatabil.

Pentru fiecare modalitate j a variabilei nominale T^q se definesc variabilele auxiliare $z_{ij,q}$: $z_{ij,q} = 1$ dacă $m_{iq} \neq 0$ și $z_{ij,q} = 0$ în rest. Matricea $\boldsymbol{Z_q}=\{z_{ij,q}\}^{n}_{i=1}{}^{\tau_q}{}_{j=1} \in \mathcal{M}_{n\times\tau q}(\mathcal{R})$, $(\forall)q \in [1, s]$, în care fiecare linie conține $\tau_q - 1$ zerouri și un singur unu, se numește matrice auxiliară a modalităților variabilei nominale T^q. Matricea $\boldsymbol{Z} = [\boldsymbol{Z_1} \vdots \cdots \vdots \boldsymbol{Z_q} \vdots \cdots \vdots \boldsymbol{Z_s}] \in \mathcal{M}_{n\times p}(\mathcal{R})$, obținută prin concatenarea matricilor $\boldsymbol{Z_q}$, se numește tabel disjunctiv complet iar $\boldsymbol{B} = \boldsymbol{Z'Z} \in \mathcal{M}_{p\times p}(\mathcal{R})$, se numește tabel de contingență Burt asociat tabelului disjunctiv complet $\boldsymbol{Z}$.

Analiza corespondențelor multiple este analiza corespondențelor simple aplicată unui tabel disjunctiv complet. Proximitatea între indivizi semnifică faptul că au ales global aceleași modalități, proximitatea între modalități semnifică faptul că ele, fie au fost alese de grupe de indivizi asemănători, fie că grupele de indivizi care le-au ales sunt asemănătoare. Regulile de interpretare a rezultatelor privind elementele active ale unei analize a corespondențelor multiple sunt asemănătoare cu cele corespunzătoare unei analize a corespondențelor simple.

În *analiza canonică* sunt explorate relațiile ce pot exista între două grupuri distincte de variabile cantitative, observate pe aceeași mulțime de indivizi, pentru a vedea dacă acestea descriu același fenomen, caz în care prospectorul de date ar putea renunța la unul din ele.

Observațiile disponibile sunt descrise în două matrici: $\boldsymbol{X'} = \{x'_{ij'}\}^{n}_{i=1}{}^{p'}{}_{j'=1} \in \mathcal{M}_{n\times p'}(\mathcal{R})$ și $\boldsymbol{X''} = \{x''_{ij''}\}^{n}_{i=1}{}^{p''}{}_{j''=1} \in \mathcal{M}_{n\times p''}(\mathcal{R})$, unde n reprezintă numărul de indivizi, p' (respectiv, p'') reprezintă numărul de variabile din primul (respectiv, al doilea) grup iar elementul $x'_{ij'}$ (respectiv, $x''_{ij''}$) reprezintă valoarea variabilei j' (respectiv, j'') măsurată pe individul i. În spațiul $\mathcal{F}$ al variabilelor, respectiv $\mathcal{R}^n$ înzestrat cu o bază canonică $\mathcal{F}$ și cu o metrică $\boldsymbol{M}$, se pot defini două subspații vectoriale: $\mathcal{F}_{X'}$, generat de vectorii-coloană $\{\boldsymbol{x'_{j'}}\}^{p'}{}_{j'=1}$, în general de dimensiune p' și $\mathcal{F}_{X''}$, generat de vectorii-coloană $\{\boldsymbol{x''_{j''}}\}^{p''}{}_{j''=1}$, în general de dimensiune p'', de asemenea, pentru indivizi pot fi luate în considerație două spații vectoriale: $\mathcal{E}_1 = (\mathcal{R}^p, \mathcal{E}, \boldsymbol{M})$, generat de vectorii-linie $\{\boldsymbol{x'_i}\}^{n}_{i=1}$ și $\mathcal{E}_2 = (\mathcal{R}^p, \mathcal{E}, \boldsymbol{M})$, generat de vectorii-linie $\{\boldsymbol{x''_i}\}^{n}_{i=1}$.

Considerând matricile $\boldsymbol{P_{X'}}$ și $\boldsymbol{P_{X''}}$ (matricile proiecțiilor ortogonale ale lui $\mathcal{F}$ înzestrat cu metrica $\boldsymbol{I}$ pe subspațiile $\mathcal{F}_{X'}$ și respectiv $\mathcal{F}_{X''}$) se obține un număr de p cupluri $\{(V^s, W^s)\}^{p}_{s=1}$ de variabile canonice care țin cont de legăturile liniare dintre cele două grupe de variabile inițiale și în care: vectorii V^s sunt vectorii proprii normați ai matricii $\boldsymbol{P_{X'}P_{X''}}$ corespunzători valorilor proprii λ_s ordonate descrescător și constituie o bază ortonormată în $\mathcal{F}_{X'}$; vectorii W^s sunt vectorii proprii normați ai matricii $\boldsymbol{P_{X''}P_{X'}}$ corespunzători aceloraşi valori proprii λ_s și constituie un sistem ortonormat al lui $\mathcal{F}_{X''}$; coeficienții $\{\delta_s = \sqrt{\lambda_s}\}^{p}_{s=1}$ sunt coeficienții de corelație canonică.

Reprezentările grafice ale rezultatelor analizei canonice se fac într-o dimensiune d, redusă, $1 \leq d \leq p$ cu ajutorul vectorilor $\boldsymbol{v^s} \in \mathcal{F}_{X'}$ și $\boldsymbol{w^s} \in \mathcal{F}_{X''}$ asociați variabilelor canonice V^s și respectiv W^s. Cele două grafice (în $\mathcal{F}_{X'}$ și în $\mathcal{F}_{X''}$) având aceeași calitate și conducând la aceleași interpretări este suficient unul singur pentru a interpreta rezultatele unei analize. În măsura în care graficul astfel obținut este „bun” el poate fi utilizat pentru a interpreta relațiile (proximități, opoziții, depărtări) dintre cele două mulțimi de variabile. În fiecare din spațiile relative la indivizi ($\mathcal{E}_1$ și $\mathcal{E}_2$) se poate, deasemenea, obține câte o reprezentare grafică a acestor indivizi în dimensiunea d, cele două reprezentări fiind comparabile, cu atât mai comparabile cu cât corelațiile canonice sunt mai mai ridicate. Coordonatele indivizilor pe axele canonice în aceste două reprezentări sunt date de liniile matricilor $\boldsymbol{V_d} \in \mathcal{M}_{n\times d}(\mathcal{R})$ (în $\mathcal{E}_1$) și $\boldsymbol{W_d} \in \mathcal{M}_{n\times d}(\mathcal{R})$ (în $\mathcal{E}_2$), ale căror coloane conțin coordonatele primelor d variabile canonice, în baza canonică $\mathcal{F}$ a spațiului $\mathcal{F}$.

Analiza canonică este considerată, pe plan teoretic, una din metodele descriptive multidimensionale centrale deoarece generalizează celelalte metode dar, de asemenea, poate fi

privită ca un caz particular de analiză în componente principale a două pachete de variabile într-un spaţiu înzestrat cu o metrică specială.

Analiza grupurilor sau clasificare

Obiective. Observaţiile disponibile privesc o populaţie de n indivizi descrişi prin un număr de p variabile. Teoretic problema clasificării este simplă, mulţimea indivizilor de clasificat fiind finită, se generează toate partiţiile posibile reţinând pe aceea (acelea) care satisface (satisfac) un criteriu de optimalitate dat. Această abordare nu este încă realizabilă și, practic, se caută o tipologie (sau segmentare) care, prin optimizarea unui criteriu, să conducă la gruparea indivizilor în clase, fiecare clasă fiind cât mai omogenă posibil şi cât mai distinctă posibil de celelalte clase. Clasele se obţin pe baza unor algoritmi formalizaţi şi nu prin metode subiective sau vizuale ce fac apel la iniţiativa sau expertiza prospectorului de date. Obiectivul unei metode de clasificare este diferit de obiectivul metodelor de discriminare (sau clasare) pentru care tipologia este cunoscută à priori, cel puţin pentru un eşantion de învăţare. În demersul analizei grupurilor, spre deosebire de demersul analizei factoriale, compresia datelor se face procedând la reducerea numărului de indivizi, faţă de reducerea numărului de variabile. Variabilele pot fi, după caz, fie toate cantitative, fie toate binare (prezenţa sau absenţa caracteristicii), fie toate calitative, fie mixte (o parte calitative şi celelalte cantitative). Pentru oricare din situaţii se poate dispune, fie de un tabel $n \times p$ de măsuri cantitative însoţit de o matrice $p \times p$ care să definească o distanţă euclidiană, fie, direct, de un tabel $n \times n$ de disimilarități sau de distanţe între indivizi.

Abordarea ierarhică se referă la tehnica agregării după dispersie, interesantă prin compatibilitatea rezultatelor sale cu unele rezultate din analiza factorială și la tehnica saltului minimal, echivalentă dintr-un anumit punct de vedere cu căutarea arborelui minimal. Metodele de clasificare ascendentă ierarhică constau în crearea, la fiecare etapă, a unei partiţii obţinute prin agregarea celor mai apropiate două elemente (indivizi sau grupuri de indivizi deja generate). Metodele nu furnizează o partiţie în q clase a unei mulţimi de n obiecte ci o ierarhie de $n - 1$ partiţii sub forma unui arbore (dendogramă). Cunoscând arborele de clasificare este ușor să se obţină o partiţie cu un număr mai mic sau mai mare de clase, pentru aceasta este suficient să se „taie” arborele la un nivel dat şi să se considere clasele furnizate de ramurile care se desprind. Fiecare „tăiere” a arborelui determină o partiţie având cu atât mai puţine clase, şi acestea fiind cu atât mai puţin omogene, cu cât tăierea se face mai sus. Interesul pentru acest arbore este dat de faptul că acesta poate oferi o idee privind numărul de clase ce există efectiv în populaţie.

Notând cu $\boldsymbol{E}$ mulţimea (finită) a indivizilor, o mulţime de mulţimi $\mathcal{H} \subseteq \mathcal{P}(\boldsymbol{E})$ se numeşte ierarhie, dacă şi numai dacă $\boldsymbol{E}$ aparţine lui $\mathcal{H}$, părţile mulțimii $\boldsymbol{E}$ formate dintr-un singur element aparţin lui $\mathcal{H}$ și $\boldsymbol{A} \cap \boldsymbol{B} \in \{\boldsymbol{A}, \boldsymbol{B}, \varnothing\}$, $(\forall)\boldsymbol{A}, \boldsymbol{B} \in \mathcal{H}$. Elementele din $\mathcal{H}$ se numesc partiţii ale lui $\boldsymbol{E}$, elementele unei partiţii se numesc clase, fiecărei ierarhii îi corespunde un arbore de clasificare, fiecare clasă dintr-o ierarhie este reuniunea claselor incluse în ea. Dacă $card(\boldsymbol{E}) = n$ atunci $card(\mathcal{H}) = n$; partiţia P_n , cu n clase, este formată din elementele mulţimii $\boldsymbol{E}$ şi conţine câte un singur element în fiecare clasă; partiţia P_1, cu o clasă, este formată din mulţimea $\boldsymbol{E}$. Se definește indicele unei ierarhii $\mathcal{H}$ ca fiind o aplicaţie, $i : \mathcal{H} \to \mathcal{R}_+$, crescătoare, adică $[(\forall)\boldsymbol{A}, \boldsymbol{B} \in \mathcal{H}, \boldsymbol{A} \subset \boldsymbol{B})] \Rightarrow [i(\boldsymbol{A}) < i(\boldsymbol{B})]$, care îndeplineşte condiţia $i(\boldsymbol{C}) = 0$, $(\forall)\boldsymbol{C} \in P_n$. Indicele i al ierarhiei $\mathcal{H}$, dacă există, se mai numeşte și nivel de agregare iar ierarhia $\mathcal{H}$ se numeşte ierarhie indexată. Dacă $\delta : \boldsymbol{E} \times \boldsymbol{E} \to \mathcal{R}_+$ este o disimilaritate strictă pe $\boldsymbol{E}$ atunci

indicele i definit prin 0 dacă $\boldsymbol{A}=\{i\}$, $i\in\boldsymbol{E}$ sau $min\ \delta(i, j)$ dacă $\boldsymbol{A}=\boldsymbol{A_1}\cup\boldsymbol{A_2}$, $\boldsymbol{A_1}\cap\boldsymbol{A_2}=\emptyset$, $i\in\boldsymbol{A_1}$, $j\in\boldsymbol{A_2}$ induce pe $\boldsymbol{E}$ o ierarhie indexată cu nivelul de agregare i.

In funcţie de natura spaţiului în care se găsesc indivizii de agregat, pentru construcţia arborelui de clasificare se pot folosi: metoda Ward, dacă indivizii formează un nor de puncte într-un spaţiu euclidian (de exemplu $\mathscr{R}^p$) unde între ei se poate calcula o distanţă euclidiană sau strategii de agregare pe disimilarităţi, dacă între indivizi se poate calcula o disimilaritate strictă.

Pe baza distanţei euclidiene se poate evalua inerţia şi astfel se poate utiliza principiul de agregare ce reuneşte acele clase pentru care inerţia interclase descreşte cel mai puţin. Conform principiului lui Huygens, inerţia globală este suma inerţiilor interclase şi intraclase. Cu cât clasele sunt mai omogene cu atât inerţia intraclase este mai mică, deci inerţia interclase este mai mare. Clase omogene înseamnă clase cu indivizi cât mai puţini, deci partiţii cât mai bogate. Este firesc ca, prin fuzionarea a două clase, inerţia intraclase să crească, deci inerţia interclase să scadă. Se va alege, deci, acea fuzionare pentru care inerţia interclase scade cel mai puţin, adică sunt grupate clasele cele mai asemănătoare (cele mai apropiate). Pierderea de inerţie interclase este $\delta(\boldsymbol{A}, \boldsymbol{B}) = P_A P_B d^2(\boldsymbol{g_A}, \boldsymbol{g_B}) / (P_A + P_B)$, unde $\boldsymbol{A}$ şi $\boldsymbol{B}$ sunt două clase cu ponderile P_A, P_B şi centrele de greutate $\boldsymbol{g_A}$, $\boldsymbol{g_B}$ sau, (confom formulei Lance-Williams generalizate) $\delta(\boldsymbol{C},(\boldsymbol{A}, \boldsymbol{B})) = ((P_A + P_C)\delta(\boldsymbol{A}, \boldsymbol{C}) + (P_B + P_C)\delta(\boldsymbol{B}, \boldsymbol{C}) - P_C\,\delta(\boldsymbol{A}, \boldsymbol{B}))/(P_A + P_B + P_C)$. Intr-o ierarhie indexată, agregată pe baza unei distanţe euclidiene, suma indicilor de agregare este egală cu inerţia totală.

Proprietăţile de mai sus permit calculul disimilarităţii dintre două clase fără a fi necesară folosirea distanţelor euclidiene între centrele de greutate al acestor clase. In plus, nici centrele de greutate nu mai trebuie calculate. Odată calculate disimilarităţile dintre indivizi, se poate lucra numai pe matrici de disimilarităţi prin aplicarea succesivă a formulei Lance-Williams. Dacă între indivizi există dată o matrice de disimilaritate strictă, atunci se pot imagina mai multe soluţii, dintre care cele mai utilizate sunt: distanţa saltului minimal (*single linkage*), care favorizează mulţimile cu puncte apropiate $d(\boldsymbol{A}, \boldsymbol{B}) = min_{(i, j)}\delta(\boldsymbol{e_i}, \boldsymbol{e_j})$, $\boldsymbol{e_i}\in\boldsymbol{A}$, $\boldsymbol{e_j}\in\boldsymbol{B}$; distanţa diametrului (*complete linkage*), ce atenuează limitele primei distanţe dar punctele trebuie să fie apropiate $d(\boldsymbol{A}, \boldsymbol{B}) = max_{(i, j)}\ \delta(\boldsymbol{e_i}, \boldsymbol{e_j})$, $\boldsymbol{e_i}\in\boldsymbol{A}$, $\boldsymbol{e_j}\in\boldsymbol{B}$ şi distanţa mediei (*unweighted pair-group average linkage*) $d(\boldsymbol{A}, \boldsymbol{B}) = P_x\delta(x, z) + P_y\delta(y, z)$ cu $\boldsymbol{A} = \{x, y\}$, $\boldsymbol{B} = \{z\}$.

Ierarhiile induse de diferitele distanţe sunt în general diferite. Pentru prospectorul de date se recomandă utilizarea mai multor tipuri de clasificări. Acestea nu trebuie să difere prea mult când se priveşte partea superioară a arborelui de clasificare. Dacă totuşi acest lucru se întâmplă, se poate conchide că mulţimea indivizilor se pretează prost la orice clasificare.

Abordarea neierarhică procedează la căutarea directă a unei partiţii şi se referă la metodele de agregare în jurul centrelor mobile, înrudite cu metoda norilor dinamici sau cu metoda celor k-medii, metode gratifiante în cazul tabelelor mari. Scopul fiecărei clasificări fiind acela de a obţine clase cât mai omogene, iar omogenitatea fiind caracterizată, din punct de vedere statistic, de dispersie, rezultă că o clasă va fi cu atât mai omogenă cu cât inerţia norului de puncte ce o alcătuieşte este mai mică. Metodele de clasificare neierarhică permit clasificarea rapidă, a unor mulţimi destul de mari de indivizi, optimizând local un criteriu de tip inerţie, criteriu care presupune cunoaşterea à priori a numărului de clase. Compararea a două partiţii cu număr diferit de clase nu este posibilă deoarece cea mai bună partiţie de k clase va avea o inerţie intraclase superioară oricărei partiţii de $k + 1$ clase, iar la limită, cea mai bună partiţie este cea trivială în care fiecare individ formează o clasă.

Se doreşte clasificarea unei mulțimi $\boldsymbol{E}$ de n indivizi caracterizaţi de p variabile în k clase, unde k este cunoscut à priori. Spaţiul $\mathscr{R}^p$, ce conţine norul de n puncte-indivizi, se presupune că este dotat cu o distanţă d corespunzătoare (distanţa euclidiană uzuală sau distanţa χ^2).

Pentru metoda centrelor mobile se prezintă un algoritm iterativ care pornește prin alegerea, în general aleator, a k puncte distincte (centre) din $\boldsymbol{E}$, $\boldsymbol{C} = \{\boldsymbol{c}_\ell\}^k_{\ell=1} \subset \boldsymbol{E}$. În fiecare iterație j se determină: distanțele dintre centrele $\boldsymbol{c}_\ell$ și elementele lui $\boldsymbol{E}$, $\boldsymbol{D} = \{ d(\boldsymbol{e}_i, \boldsymbol{c}_\ell) \}^n_{i=1}{}^k_{\ell=1}$; clasele cu centrele $\boldsymbol{c}_\ell$, $\boldsymbol{E}_{c\ell} = \{\boldsymbol{e}_i \in \boldsymbol{E} \mid d(\boldsymbol{e}_i, \boldsymbol{c}_\ell) \le d(\boldsymbol{e}_i, \boldsymbol{c}_{\ell'}),\ \ell' = 1 \div k,\ \ell' \neq \ell\}$; centrele de greutate $\{\boldsymbol{g}_\ell\}_{\ell=1\div k}$ ale claselor $\{\boldsymbol{E}_{c\ell}\}_{\ell=1\div k}$ și inerţia intraclase $I_W^{(j+1)}$ a partiţiei $\{\boldsymbol{E}_{c\ell}\}_{\ell=1\div k}$. Dacă numărul de iterații prevăzut a fost depășit ($j > N$) sau ameliorarea inerţiei intraclase este considerată nesemnificativă ($| I_W{}^{(j+1)} - I_W{}^{(j)} | \le \varepsilon$), atunci algoritmul se oprește, altfel se trece la o nouă iterație ($j := j + 1$) luând în considerație ultimele centre de greutate calculate($\boldsymbol{c}_\ell := \boldsymbol{g}_\ell,\ \ell = 1 \div k$).

Algoritmul converge într-un număr finit de paşi, experienţa arată că viteza de convergenţă este rapidă. Trebuie remarcat şi faptul că, la fiecare pas nefiind necesar decât calculul a nk distanţe, acelea dintre cei n indivizi şi cele k centre de greutate, nu este necesară menţinerea în memorie a tabelului cu toate cele $n(n–1)/2$ distanţe dintre indivizi. Pentru a înlătura dependenţa metodei de punctele iniţiale se utilizează metoda norilor dinamici a lui E. Diday, care este o generalizare a metodei centrelor mobile în sensul că fiecare clasă nu mai este reprezentată de centrul său de greutate ci de un nucleu de puncte (cele mai centrale, de exemplu), de o axă principală şi de un plan principal.

Abordarea mixtă. Metodele de agregare ierarhice dau întotdeauna acelaşi rezultat dacă datele iniţiale sunt aceleaşi, dau indicaţii privind numărul de clase ce trebuie reţinut, dar sunt slab adaptate la volume mari de date. Metodele de agregare în jurul centrelor mobile pot manipula volume mari cu preţuri mici dar au dezavantajul că produc partiţii dependente și de numărul ales de clase şi de centrele iniţiale. Combinarea celor două metode a condus la o metodă mixtă (*hybrid clustering*). Metoda de clasificare mixtă acoperă trei aspecte: partiţionarea mulţimii elementelor de clasificat în câteva zeci (eventual sute) de partiţii omogene; obţinerea unei dendrograme care să sugereze numărul de clase finale ce trebuie reţinute; optimizarea partiţiei obţinută prin tăierea arborelui. Partiţionarea iniţială vizează obţinerea rapidă şi cu un preţ scăzut (utilizând metoda centrelor mobile) a unei partiţii de n obiecte în k clase omogene, $s \ll k \ll n$, unde s este numărul de clase dorit. Desigur, optimalitatea nu este atinsă dar partiţia obţinută poate fi ameliorată pornindu-se de la grupurile stabile. Agregarea ierarhică a claselor obţinute constă în efectuarea unei clasificări ierarhice ascendente în care elementele terminale ale arborelui sunt cele k clase ale partiţiei iniţiale. Scopul etapei este de a reconstitui clasele care au fost fragmentate şi de a agrega elementele aparent dispersate în jurul centrelor de origine. Arborele este construit prin metode de clasificare ierahică, metode care ţin seamă de mase în momentul alegerii elementelor de agregat. Partiţionarea finală a populaţiei este dată prin tăierea arborelui obţinut în etapa precedentă, omogenitatea claselor obţinute putând fi optimizată prin reafectare. Tăind arborele la nivelul unui salt important al indicelui de agregare se poate spera în obţinerea unei partiţii de bună calitate în sensul că indivizii grupaţi sub nivelul de tăiere sunt apropiaţi iar cei grupaţi deasupra nivelului de tăiere sunt necesarmente depărtaţi (ceea ce corespunde definiţiei unei bune partiţii).

Caracterizarea grupurilor. În cazul analizei grupurilor elementele unei aceleiaşi clase se aseamănă din punct de vedere al criteriilor alese pentru a le descrie și la fel ca în cazul analizei factoriale, criteriile utilizate sunt empirice. Precizarea criteriilor aflate la originea

grupurilor rezultate se obţine procedând la o descriere automată a claselor, etapă indispensabilă oricărei proceduri de clasificare. Descrierea automată a claselor este, în general, bazată pe compararea mediilor sau a procentelor din interiorul claselor cu mediile sau procentele obţinute pe întreaga populaţie. Criteriul de selecţie a variabilelor continue sau a modalităţilor variabilelor nominale, caracteristice fiecărei clase, îl constituie o valoare-test destinată să măsoare ecartul dintre valorile specifice clasei şi valorile globale. Pentru o variabila continuă, x, valoarea-test este $t_k = (\bar{x}_k - \bar{x}) / s_k(x)$, unde $s^2_k(x) = (n - n_k)s^2(x) / (n - 1)n_k$ este estimatorul dispersiei lui x în clasa k și $s^2(x)$ este dispersia empirică a lui x în întreg norul. Pentru modalitatea j valoarea-test (sau abundenţa) este definită comparând procentul ei în clasă, n_{jk} / n_k , cu procentul ei în toată populaţia, n_j / n unde, n_{jk} reprezintă numărul de indivizi având modalitatea j dintre cei n_k indivizi ai clasei k și n_j reprezintă numărul de indivizi având modalitatea j dintre toți cei n indivizi.

Metode explicative

Modelare în vederea previziunii

Instruire. Având la dispoziţie o serie de observaţii asupra unei variabile p-dimensionale X (mulțimea variabilelor explicative, $X = \{X^j\}^p_{j=1}$) măsurată pe o mulţime de n indivizi, în funcţie de prezenţa sau absenţa unei variabile de explicat Y, observată în conjuncţie cu X, se pot distinge două tipuri de probleme, numite de instruire: în prezenţa variabilei de explicat Y este vorba de o problemă de instruire supervizată sau de modelare „să se găsească o funcţie φ susceptibilă să reproducă cel mai bine pe Y, conform unui criteriu de definit, observându-l pe X, $Y = \varphi(X) + \varepsilon$, unde ε simbolizează eroarea de măsurare sau zgomotul”; în absenţa variabilei de explicat este vorba de o problemă de instruire nesupervizată: „să se găsească o tipologie sau taxinomie a observaţiilor, cum să fie acestea regrupate în clase cât mai omogene dar cât mai diferite intre ele”. În demersul inferenţial şi confirmatoriu obiectivul principal urmărit îl constituie modelarea sau discriminarea respectiv deducerea unui model de previziune pentru variabila ţintă. Modelele și metodele cele mai frecvent utilizate în atingerea acestui obiectiv sunt: modelele liniare, metodele de discriminare (geometrice și probabiliste), metodele conexioniste, maşinile cu suport vectorial, , metodele de segmentare, metodele de agregarea a modelelor (*Bagging*, *Random Forest*, *Boosting*).

Calitatea previziunii. Performanța unui model, rezultat al unei metode de instruire, se evaluează prin capacitatea sa de previziune sau de generalizare. Măsurarea acestei performanțe este foarte importantă pentru prospectorul de date deoarece permite selecția unui model optim dintr-o familie de modele asociată metodei de învățare utilizate, ghidează alegerea metodei comparând modelele selecționate între ele și oferă o măsură a calității sau a încrederii care se poate acorda previziunii. Estimarea calității previziunii este un element central al oricărei strategii de *data mining*. În principiu, sunt avute în vedere trei tipuri de abordări: partiționarea eșantionului pentru a separa estimarea modelului de estimările erorii de previziune, penalizarea erorii de ajustare luând în cosiderație complexitatea modelului sau recurgerea la simulări implicând multiplicarea calculelor. Alegerea oricărei abordări depinde de mai mulți factori între care dimensiunea eșantionului inițial, complexitatea modelului anvizajat, varianța erorii, complexitatea algoritmilor adică volumul de calcule admisibil.

Dacă, F reprezintă legea lui Y în conjuncție cu X, $\mathbf{z} = \{(\mathbf{x}_i, y_i)\}^n_{i=1}$ reprezintă un eșantion, X și fiecare $\mathbf{x}_i$ fiind de dimensiune p iar $Y = \varphi(X) + \varepsilon$ reprezintă modelul de estimat, cu ε independent de X, $E(\varepsilon) = 0$ și $var(\varepsilon) = \sigma^2$, atunci eroarea de previziune a modelului poate fi

definită prin $\mathcal{E}_P(\mathbf{z}, F) = E_F[Q(Y, \tilde{\varphi}(X))]$, unde Q este o funcție de pierdere. Dacă variabila Y de previzionat este cantitativă funcția de pierdere este, în general, pătratică $Q(y, \tilde{y}) = (y - \tilde{y})^2$ iar dacă Y este calitativă Q este un indicator de misclasare $Q(y, \tilde{y}) = 1_{\{y \neq \tilde{y}\}}$. In cazul cantitativ eroarea de previziune, într-un punct $\boldsymbol{x}$, se descompune astfel $\mathcal{E}_P(\boldsymbol{x}) = \sigma^2 + bias^2 +$ varianță. Cu cât un model este mai complex, adică cu un număr mai mare de parametri, cu atât el este mai flexibil, respectiv, se poate ajusta cu atât mai bine la datele observate și deci *bias*-ul său va putea fi cu atât mai redus. Dar, pe de altă parte, varianța crește odată cu numărul de parametri de estimat adică odată cu complexitatea modelului. Pentru a minimiza riscul pătratic definit mai sus, soluția este de a căuta un compromis cât mai bun între *bias* și varianță, de a accepta *bias*-area estimării pentru a reduce cât mai favorabil varianța.

Un crieriu de estimare a erorii de previziune, care exprimă calitatea de ajustare a modelului pe eșantionul observat, este $\mathcal{E}_P = 1/n \sum_{i=1}^{n} Q(y_i, \tilde{\varphi}(\boldsymbol{x}_i))$. În cazul cantitativ, acest criteriu este minimizat prin cercetarea celor mai mici pătrate, în cazul calitativ estimarea este rata de misclasare. Modul cel mai simplu de a estima, fără *bias*, eroarea de previziune constă în a calcula $\mathcal{E}_P$ pe un eșantion independent care nu a participat la estimarea modelului. Dacă dimensiunea eșantionului este suficient de mare, se procedează la separarea eșantionului în trei părți numite respectiv de învățare, de validare și de test ($\mathbf{z} = \mathbf{z}_{\hat{i}nv} \cup \mathbf{z}_{val} \cup \mathbf{z}_{test}$), $\tilde{\mathcal{E}}_P(\mathbf{z}_{\hat{i}nv})$ este minimizată pentru a estima un model; $\tilde{\mathcal{E}}_P(\mathbf{z}_{val})$ servește la compararea modelelor în interiorul unei aceleiași familii pentru a-l selecționa pe acela care minimizează această eroare; $\tilde{\mathcal{E}}_P(\mathbf{z}_{test})$ este utilizată pentru a compara între ele cele mai bune modele ale fiecărei metode considerate. Dacă dimensiunea eșantionului este insuficientă calitatea ajustării este degradată, varianța estimării erorii poate fi importantă dar nu poate fi estimată și atunci selecția modelului se bazează pe un alt tip de estimare a erorii de previziune recurgându-se, fie la o penalizare, fie la simulări (validare încrucișată).

Bootstrap. Motivul pentru care se recurge la tehnicile de *bootstrap* (sau re-eșantionare) îl constituie evaluarea, prin simulare, a distribuției unui estimator atunci când nu se cunoaște legea eșantionului sau, de cele mai multe ori, atunci când nu se poate presupune că este gaussiană. Obiectivul este de a înlocui ipotezele probabilistice, nu totdeauna verificate sau chiar neverificabile, prin simulări implicând mai multe calcule. Ideea de bază a *bootstrap* constă în substituirea distribuției de probabilitate F, necunoscută, aferentă eșantionului de învățare, cu distribuția empirică $\tilde{F}$ obținută acordând o pondere de $1/n$ fiecărei realizări. Astfel se obține un eșantion de dimensiune n numit eșantion *bootstrap* cu legea de distribuție empirică $\tilde{F}$ prin n extrageri aleatoare cu înlocuire dintre cele n observații inițiale. Este comod să se construiască un număr mare de eșantioane *bootstrap* pe care să se calculeze estimatorul respectiv. Legea simulată a acestui estimator este o aproximare asimptotic convergentă, în ipoteze rezonabile, a legii estimatorului. Această aproximare oferă estimări ale *bias*-ului, ale varianței, deci a unui risc pătratic, și chiar intervalele de încredere ale estimatorului, fără vre-o ipoteză (normalitate) privind legea reală.

Fie $\boldsymbol{z}$ un eșantion *bootstrap* al datelor: $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}^{\boldsymbol{n}}_{i=1}$. Estimatorul *plug-in* al erorii de previziune, $\mathcal{E}_P(\mathbf{z}, F)$, pentru care distribuția F este înlocuită cu distribuția empirică $\tilde{F}$, este definit prin: $\mathcal{E}_P(\boldsymbol{z}, \tilde{F}) = (1/n) \sum_{i=1}^{n} nQ(y_i, \varphi_z(\boldsymbol{x}_i))$, unde φ_z reprezintă estimarea lui φ pe eșantionul *bootstrap*. Estimarea *bootstrap* a erorii medii de previziune, $E_F[\mathcal{E}_P(\mathbf{z}, F)]$, este dată de: $\mathcal{E}_{boot} = E_{\tilde{F}}[\mathcal{E}_P(\boldsymbol{z}, \tilde{F})] = E_{\tilde{F}}[(1/n) \sum_{i=1}^{n} nQ(y_i, \varphi_z(\boldsymbol{x}_i))]$, iar estimarea obținută prin simulare va fi: $\tilde{\mathcal{E}}_{boot} = (1/K) \sum_{\kappa=1}^{K} (1/n) \sum_{i=1}^{n} nQ(y_i, \varphi_{z\kappa}(\boldsymbol{x}_i))$. Estimarea erorii de previziune astfel construită este, în general, *bias*-ată prin optimism deoarece, datorită simulărilor, aceleași observații apar în același timp și în estimarea modelului și în estimarea erorii. Există

abordări care vizează corecția acestui *bias*. Estimatorul *out-of-bag* al erorii de previziune $\tilde{\mathcal{E}}_{oob}$, inspirat din validarea încrucișată, consideră, pe de o parte, observațiile extrase în eșantionul *bootstrap* și, pe de altă parte, observațiile neutilizate la estimarea modelului dar reținute pentru estimarea erorii: $\tilde{\mathcal{E}}_{oob} = (1/n)\sum^{n}_{i=1} 1/B_i \sum_{\kappa \in \mathcal{K}i} Q(y_i, \varphi_{z\kappa}(\boldsymbol{x}_i))$, unde $\mathcal{K}_i$ reprezintă mulțimea de indici κ ai eșantioanelor *bootstrap* neconținând a *i*-a observație după cele *B* simulări și B_i reprezintă numărul $|\mathcal{K}_i|$ al acestor eșantioane. *B* trebuie să fie suficient de mare pentru ca orice observație să poată să fie extrasă cel puțin o dată, altfel termenii cu $\mathcal{K}_i = 0$ trebuiesc omiși.

Modele liniare

Modelele liniare urmăresc să prevadă (să explice sau să prezică) o variabilă continuă, numită variabilă de explicat (dependentă sau endogenă) cu ajutorul unor variabile numite explicative (exogene sau predictori). În cazul în care variabilele explicative sunt continue modelul este un model de analiză a regresiei, dacă acestea sunt variabile discrete (nominale) modelul este de analiză dispersională (sau analiză de varianţă) iar dacă mulţimea variabilelor exogene este mixtă modelul este de analiză de covarianţă.

Analiza regresiei. În modelul de analiză a regresiei relaţia dintre *Y* şi *X* este presupusă liniară, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ unde: $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$, $\boldsymbol{y} \in \mathcal{M}_{n\times 1}(\mathcal{R})$ reprezintă vectorul observaţiilor asupra variabilei dependente *Y*, $\boldsymbol{X} = \{x_{ij}, x_{i0} = 1\}^{n}_{i=1}{}^{p}_{j=0}$, $\boldsymbol{X} \in \mathcal{M}_{n\times(p+1)}(\mathcal{R})$ este matricea observaţiilor asupra variabilelor explicative, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$, $\boldsymbol{\beta} \in \mathcal{M}_{(p+1)\times 1}(\mathcal{R})$ reprezintă vectorul coeficienţilor iar $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$, $\boldsymbol{\varepsilon} \in \mathcal{M}_{n\times 1}(\mathcal{R})$ este vectorul erorilor/reziduurilor. Pentru evaluarea coeficienţilor necunoscuţi ai modelului, inclusiv a reziduurilor ε_i, se dispune de un sistem de *n* ecuaţii liniare având *n+p+1* necunoscute. Sistemul admite o infinitate de soluţii; o soluţie posibilă $\boldsymbol{b} = (b_0, b_1, \ldots, b_p)$ va trebui să minimizeze global mulţimea distanţelor la modelul liniar urmând un anumit criteriu; sunt aleși acei vectori $\boldsymbol{b}$ care minimizează mulţimea valorilor $\{e_i\}^{n}_{i=1}$, unde $e_i = y_i - (b_0 + b_1 x_{i1} + \ldots + b_p x_{ip})$. Criteriul celor mai mici pătrate conduce la calcule algebrice simple, se pretează la interpretări geometrice clare şi permite interpretări interesante, motiv pentru care se utilizează cel mai des. Estimarea funcţiei de regresie liniară multiplă presupune determinarea tuturor coeficienţilor $b_0, b_1, \ldots, b_p$ prin metoda celor mai mici pătrate pornind de la observaţiile $\{ y_i, x_{i0} = 1, x_{i1}, \ldots, x_{ip} \}^{n}_{i=1}$. Se presupune că variabilele sunt centrate, ceea ce implică $b_0 = 0$; coeficienţii funcţiei de regresie liniară multiplă sunt $\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$. Căutarea lui $\boldsymbol{y}$ sub forma unei combinaţii liniare de $\boldsymbol{x}_i$ se reduce la a defini $\tilde{\boldsymbol{y}}$ într-un subspaţiu V_X generat de variabilele explicative. Metoda ajustării celor mai mici pătrate se reduce la aproximarea lui $\boldsymbol{y}$ prin proiecţia sa ortogonală $\tilde{\boldsymbol{y}}$, pe V_X înlocuindu-l pe $\boldsymbol{b}$. Se obţine $\tilde{\boldsymbol{y}} = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = P_X \boldsymbol{y}$, unde $P_X = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$, este operatorul proiecţiei ortogonale pe V_X. Lungimile în $\mathcal{R}^n$ pot fi interpretate în termeni de dispersie deoarece $(1/n)\sum^{n}_{i=1} y_i^2 = (1/n)\sum^{n}_{i=1}(y_i - \tilde{y})^2 + (1/n)\sum^{n}_{i=1}\tilde{y}_i^2$, unde $(1/n)\sum^{n}_{i=1} y_i^2$, este dispersia totală, $(1/n)\sum^{n}_{i=1}(y_i - \tilde{y})^2$, este dispersia reziduală și $(1/n)\sum^{n}_{i=1}\tilde{y}_i^2$ reprezintă dispersia explicată (a modelului).

Pentru o evaluare globală a calităţii aproximării se definesc coeficientul de corelaţie multiplă, $R = cor(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = cor(\boldsymbol{y}, \boldsymbol{Xb})$ și coeficientul de determinare $R^2 = \Sigma^{n}_{i=1}\tilde{y}_i^2 / \Sigma^{n}_{i=1} y_i^2$ (adică dispersia explicată împărțită la dispersia totală) $= \boldsymbol{y'X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} / \boldsymbol{y'y}$ (în funcţie de datele iniţiale). Dacă $R^2 = 1$ atunci $\tilde{y}_i = y_i \ (\forall) i = 1 \div n$ adică modelul liniar ajustează perfect datele.

Prin minimizarea termenului $\sum^{n}_{i=1} e_i^2$ se maximizează termenul R^2, cu alte cuvinte metoda celor mai mici pătrate determină acea combinaţie liniară a variabilelor explicative ce maximizează corelaţia cu variabila explicată $\boldsymbol{y}$.

Din punctul de vedere al prospectorului de date aspectele cele mai interesante privesc semnificaţiile statistice ale coeficienţilor de regresie, adecvarea modelului regresiei multiple la datele observate, studiul reziduurilor, observațiilor aberante și influenţei observaţiilor asupra rezultatelor, stabilizarea coeficienţilor de regresie și tehnicile de obţinere de coeficienţi stabili precum și metodele de selecţie a variabilelor ($\boldsymbol{y}$ se „explică” doar prin $q \ll p$ predictori) pentru a micşora numărul de predictori, a creşte viteza de calcul şi a obţine formule stabile cu o putere predictivă bună.

Analiza dispersională. Dacă variabilele explicative sunt discrete (nominale) regresia multiplă devine analiză dispersională. Se dispune în acest caz de n observaţii asupra variabilei continue Y observată în conjuncție cu cele p variabile nominale $\{X^k\}_{k=1}^{p}$ având, respectiv, modalităţile $\{\tau_k\}_{k=1}^{p}$.

Matricea variabilelor explicative, $\boldsymbol{X}$, se prezintă sub forma unui tablou disjunctiv complet $[\boldsymbol{X_1} \vdots \cdots \vdots \boldsymbol{X_k} \vdots \cdots \vdots \boldsymbol{X_p}]$. Pentru fiecare submatrice $\boldsymbol{X_k}$ suma coloanelor este egală cu vectorul $\boldsymbol{1_n}$ existând p relaţii liniare între coloanele lui $\boldsymbol{X}$. Sistemul de ecuaţii normale are o infinitate de soluţii, toate soluţiile duc la acelaşi vector $\tilde{\boldsymbol{y}}$ care este proiecţia lui $\boldsymbol{y}$ pe V_X, dar coeficienţii $\boldsymbol{b}$ nu sunt unici. Pentru a obţine o estimaţie unică b, trebuie impuse p restricţii liniare privind codificările variabilelor calitative. Cea mai des utilizată restricţie este ca suma coeficienţilor lui $\boldsymbol{b}$, relativ la fiecare variabilă nominală, să fie nulă, aceasta revine la suprimarea unei coloane din fiecare submatrice şi la înlocuirea coloanelor rămase cu diferenţa dintre ele şi coloana suprimată. Matricea $\ddot{\boldsymbol{X}}$, a variabilelor explicative astfel recodate este de rang maxim $rang(\ddot{\boldsymbol{X}}) = \sum_{k=1}^{p}(m_k - 1)$. Pentru exemplificare, în cazul în care se dispune de două variabile nominale A şi B, numite factori, având I respectiv J modalităţi, numite nivele, analiza dispersională cu doi factori cu interacţiune se reduce la a efectua regresia lui $\boldsymbol{y}$ cu matricea de condiţie $\ddot{\boldsymbol{X}} = [\boldsymbol{1} \vdots \ddot{\boldsymbol{X}}_1 \vdots \ddot{\boldsymbol{X}}_2 \vdots \ddot{\boldsymbol{X}}_{12}]$ cu $rang(\ddot{\boldsymbol{X}}_1)=J$; $rang(\ddot{\boldsymbol{X}}_2)=K$; $rang(\ddot{\boldsymbol{X}}_{12})=JK$, unde $\ddot{\boldsymbol{X}}_1$ şi $\ddot{\boldsymbol{X}}_2$ sunt matricile indicator reduse ale celor doi factori A şi B iar $\ddot{\boldsymbol{X}}_{12}$ este matricea interacţiunilor corespunzând celor JK combinaţii ale nivelelor lui A şi B. In această situaţie modelul liniar devine $\boldsymbol{y} = \boldsymbol{\mu 1} + \ddot{\boldsymbol{X}}_1\boldsymbol{\alpha} + \ddot{\boldsymbol{X}}_2\boldsymbol{\beta} + \ddot{\boldsymbol{X}}_{12}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ și deci se poate utiliza un program de regresie multiplă pentru a efectua o analiză dispersională. Procedura poate fi generalizată la modele cu mai mulţi factori şi nivele de interacţiune de ordin superior.

Totuşi, o anumită prudenţă se impune din mai multe motive: este dificil de apreciat şi de limitat clar natura ipotezelor testate; interacţiunile de ordin superior pot duce la „teste în lanţ” delicat de interpretat; o interacţiune, mai ales de ordin superior, se poate datora prezenţei unor observaţii uşor aberante, caz în care procedura nu este robustă.

Modelele liniare generalizate extind modelele liniare clasice în două direcţii: combinaţia liniară $a_i = b_0x_{i0} + b_1x_{i1} + ... + b_px_{ip}$ a variabilelor explicative poate fi o funcţie l de $E(y_i)$ (numită funcţie de legătură) adică $a_i = l(E(y_i))$ în comparaţie cu modelele liniare obișnuite în care $a_i = E(y_i)$; legea de probabilitate a lui $\boldsymbol{y}$ poate fi şi un alt membru al clasei legilor exponenţiale (legile binomiale, Poisson, Gamma) decât legea normală. Alegând diferite legi de probabilitate din clasa legilor exponenţiale şi diferite funcţii de legătură pentru $\boldsymbol{y}$, se pot obţine şi alte modele, printre care un loc important îl ocupă modelele log-liniare.

Ajustarea modelelor liniare generalizate se face prin metoda verosimilităţii maxime care, în cazul legii normale, coincide cu metoda celor mai mici pătrate.

Metode de discriminare.

Metode geometrice. Metodele geometrice de analiză discriminantă, esenţialmente descriptive, se bazează pe noţiunea de distanţă şi nu utilizează nici o noţiune probabilistă.

Se dispune de observaţii privind p variabile cantitative $\{X^j\}_{j=1}^{p}$, jucând rolul de variabile explicative şi o variabilă calitativă Y cu q modalităţi $\{k\}_{k=1}^{q}$, jucând rolul de variabilă de explicat. Cele p variabile explicative X^j au fost observate pe un eşantion $\{\boldsymbol{x}_i\}_{i=1}^{n}$, de n indivizi. Variabila nominală Y generează o partiţie a celor n indivizi în q clase $\{A_k\}_{k=1}^{q}$.

Problema de discriminare (sau clasare) este următoarea: „fiind dat un nou individ $\boldsymbol{x}$ pe care au fost observate variabilele explicative X^j dar nu și variabila de explicat Y se pune problema de a decide modalitatea k a lui Y (sau clasa A_k corespunzătoare) pentru $\boldsymbol{x}$ ”.

In context geometric, discriminarea poate fi interpretată ca o împărţire a spaţiului indivizilor în regiuni, $\mathcal{R}$, numite regiuni de decizie, fiecare regiune fiind asociată cu o clasă de indivizi. Regiunile de decizie şi implicit clasele corespunzătoare, se zic separabile dacă pot fi separate prin suprafeţe, S, numite suprafeţe de decizie. Dacă suprafeţele de decizie sunt hiperplane $\mathcal{H}$, clasele se zic liniar separabile. Suprafeţele de decizie pot fi descrise cu ajutorul unei mulţimi $\mathcal{G} = \{g\}$ de funcţii numite funcţii de discriminare sau funcţii de decizie. Funcţia de discriminare g ataşează fiecare individ $\boldsymbol{x}$ unei regiuni $\mathcal{R}$, regiune delimitată prin intermediul unei mulţimi de suprafeţe de decizie. Funcţia de discriminare este instruită întro fază de instruire când sunt stabilite clasele şi suprafeţele de decizie. În faza de lucru (sau decizională sau de afectare) funcţiei de discriminare i se prezintă date ale căror clase nu se cunosc, noii indivizi fiind asociați uneia sau alteia dintre clasele stabilite.

Pentru rezolvarea problemelor de discriminare sunt stabilite reguli de decizie (sau de afectare) și moduri de evaluare. Se disting următoarele trei cazuri de separabilitate:

1. Fiecare clasă A_k este separată de toate celelalte printr-o singură suprafaţă de decizie. Funcţia de decizie corespunzătoare clasei A_k este $g_k(\boldsymbol{x}) : \mathcal{R}^p \to \mathcal{R}$, $k \in [1, q]$, ecuaţia suprafeţei de decizie ce separă clasa A_k de toate celelalte clase este: $g_k(\boldsymbol{x}) = 0$.

Pentru fiecare clasă A_k, $[\boldsymbol{x} \in A_k] \Rightarrow [g_k(\boldsymbol{x}) > 0]$. Pentru un punct $\boldsymbol{x}'$, nou, dacă $g_k(\boldsymbol{x}') > 0$ şi $g_\ell(\boldsymbol{x}') < 0$, $(\forall)\ell \in [1, q]$, $\ell \neq k$ atunci $\boldsymbol{x}'$ este ataşat clasei A_k. Regiunea de decizie $\mathcal{R}_k$, corespunzătoare clasei A_k, este: $\mathcal{R}_k = \{\boldsymbol{x} \in \mathcal{R}^p \mid [g_k(\boldsymbol{x}) > 0] \wedge [g_\ell(\boldsymbol{x}) < 0], (\forall)\ell \in [1, q], \ell \neq k\}$.

2. Fiecare clasă este separată de oricare alta printr-o suprafaţă de decizie. Clasele sunt două câte două separabile, cele $q(q-1)/2$ suprafeţe de decizie sunt generate de funcţiile $g_{k\ell}(\boldsymbol{x}) : \mathcal{R}^p \to \mathcal{R}$ unde $g_{k\ell}(\boldsymbol{x}) = -g_{\ell k}(\boldsymbol{x})$, $(\forall)\boldsymbol{x} \in \mathcal{R}^p$. Suprafaţa de decizie corespunzătoare claselor A_k şi A_ℓ are ecuaţia $g_{k\ell}(\boldsymbol{x}) = 0$, punctele clasei A_k se află de partea pozitivă a suprafeţei. Regula de decizie este: $\boldsymbol{x} \in A_k \Leftrightarrow g_{k\ell}(\boldsymbol{x}) > 0$ $(\forall)\ell \in [1, q]$, $\ell \neq k$. Regiunea de decizie $\mathcal{R}_k$ corespunzătoare clasei A_k este $\mathcal{R}_k = \{\boldsymbol{x} \in \mathcal{R}^p \mid g_{k\ell}(\boldsymbol{x}) > 0, (\forall)\ell \in [1, q], \ell \neq k\}$.

3. Există q funcţii de decizie. Regula de decizie este: $\boldsymbol{x} \in A_k \Leftrightarrow g_k(\boldsymbol{x}) > g_\ell(\boldsymbol{x})$, $(\forall)\ell \neq k$, $k \in [1, q]$. Regiunea de decizie $\mathcal{R}_k$ este: $\mathcal{R}_k = \{\boldsymbol{x} \in \mathcal{R}^p \mid g_k(\boldsymbol{x}) > g_\ell(\boldsymbol{x}), (\forall)\ell \neq k\}$, $k \in [1, q]$. Suprafaţa de decizie dintre clasele A_k şi A_ℓ este dată de ecuaţia: $g_k(\boldsymbol{x}) = g_\ell(\boldsymbol{x})$, $(\forall)\boldsymbol{x} \in \mathcal{R}^p$, $(\forall)k, \ell \in [1, q]$, $\ell \neq k$. Obiectele clasei A_k se află de partea pozitivă a suprafeţei de separare.

Pentru prospectorul de date de o mare importanţă practică este cazul claselor liniar separabile. Funcțiile afine de decizie pot fi transformate în funcţii liniare de decizie. Dacă g_k este funcţia liniară de decizie corespunzând clasei A_k atunci, în conformitate cu cazul **3** de separabilitate, un obiect $\overline{\boldsymbol{x}}$ este ataşat clasei A_k dacă $g_k(\overline{\boldsymbol{x}}) > g_\ell(\overline{\boldsymbol{x}})$ $(\forall)\ell \in [1, q]$, $\ell \neq k$. În cazul **3** de separabilitate regiunile de decizie pot fi mărginite de hiperplane sau de porţiuni de hiperplane. Clasarea, prin minimizarea unei funcţii criteriu, conduce la o clasă de funcţii discriminante liniare. Funcţia criteriu luată în consideraţie este distanţa d de la vectorii caracteristică la prototipurile claselor. Un vector $\boldsymbol{x}$ este ataşat acelei clase A_k de al cărei

prototip $\boldsymbol{g}_k$ vectorul $\boldsymbol{x}$ este mai aproape, adică: $\boldsymbol{x} \in A_k$ dacă $d(\boldsymbol{x}, \boldsymbol{g}_k) = min_\ell\ d(\boldsymbol{x}, \boldsymbol{g}_\ell)$. O clasificare echivalentă se obţine considerând funcţia de decizie $\mathcal{g}_k : \mathcal{R}^p \to \mathcal{R}$ dată de formula $\mathcal{g}_k(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{g}_k - (1/2)\boldsymbol{g}'_k\boldsymbol{g}_k$. Regula de decizie devine: $\boldsymbol{x} \in A_k$ dacă $\mathcal{g}_k(\boldsymbol{x}) = max_\ell\ \mathcal{g}_\ell(\boldsymbol{x})$, $\mathcal{g}_k$ este o funcţie afină de decizie. Hiperplanul de separare este ortogonal pe dreapta ce uneşte prototipurile claselor, pe care o intersectează într-un punct situat la jumătatea distanţei dintre prototipuri. Funcţia discriminantă cu distanţă minimă este adecvată pentru cazurile când punctele unei clase tind să se aglomereze în vecinătatea unui punct prototip, formând un nor (*cluster*) de puncte.

Metode probabiliste. În abordarea probabilistă, metodele sunt dedicate aspectului inferenţial al analizei discriminante.

Fie $(\Omega, \mathcal{K}, p)$ un câmp de probabilitate. Probabilitatea condiţionată a evenimentului $A \in \mathcal{K}$ relativ la evenimentul $B \in \mathcal{K}$ cu $p(B) > 0$, este $p_B : \mathcal{K} \to \mathcal{R}$ cu $p_B(A) \equiv p(A|B) = p(A \cap B) / p(B)$.

Dacă $\{A_i\}_{i \in I} \subset \mathcal{K}$ formează un sistem complet de evenimente atunci are loc următoarea egalitate (formula lui Bayes a probabilităţii cauzelor):

$p(A_i|B) = p(A_i \cap B) / p(B) = p(A_i)p(B \cap A_i) / (p(A_i)p(B)) = p(A_i)p(B|A_i) / \sum_i p(A_i)p(B|A_i)$,

unde $\{p(A_i)\}$ sunt probabilităţile à priori şi $\{p(B|A_i)\}$ probabilităţile à posteriori. Funcţia de repartiţie a variabilei aleatoare $\mathcal{X}$ condiţionată de evenimentul $A \in \mathcal{K}$ cu $p(A) > 0$ este funcţia $F_A : \mathcal{R} \to [0, 1]$, $F_A(x) \equiv F(x \mid A) = p(X < x \mid A)$. Densitatea de repartiţie a variabilei aleatoare $\mathcal{X}$ condiţionată de evenimentul $A \in \mathcal{K}$ cu $p(A) > 0$ este funcţia $f(\bullet \mid A) : \mathcal{R} \to \mathcal{R}$, pentru care $F(x|A) = \int_{-\infty}^{x} f(t \mid A)\, dt$. $f(x|A) = F'(x|A)$ aproape peste tot. $p(A \mid \mathcal{X} = x) = p(A)f(x \mid A) / f(x)$.

Problema de discriminare (sau clasare) formulată în termenii teoriei statistice a deciziei este următoarea: „ dându-se

- m grupe (sau populaţii), $\{\Pi_k\}_{k=1}^m$, specificate prin distribuţiile lor de probabilitate, $p_k(\boldsymbol{x}) = p(\mathcal{X} = \boldsymbol{x} \mid \boldsymbol{x} \in \Pi_k)$ cu $k = 1 \div m$,
- m probabilităţi à priori, $\{q_k\}_{k=1}^m$, ca un individ (sau observaţie) să provină din populaţiile Π_k, formând un sistem complet de probabilităţi ($\sum_{k=1}^m q_k = 1$),
- $\mathcal{E} \subset \mathcal{R}^p$ spaţiul observaţiilor asupra a p variabile aleatoare, $\mathcal{X} = \{X^j\}_{j=1}^p$, (predictori),
- $\{C(j|k)\}_{k,j=1}^m$, costurile erorilor de clasare (costul clasării unui individ, provenind din populaţia Π_k, în populaţia Π_j, $j \neq k$);

să se găsească o partiţie $\mathcal{R}$ a spaţiului $\mathcal{E}$ astfel încât $\sum_{k=1}^m q_k \sum_{j=1, j \neq k}^m C(j|k)p(j|k, \mathcal{R})$ să fie minimă, unde: $\{p(j|k, \mathcal{R}) = \int_{Rj} p_k(\boldsymbol{x})d\boldsymbol{x}\}_{j=1}^m {}_{k=1, k \neq j}^m$ reprezintă probabilităţile de eroare pentru partiţia $\mathcal{R}$, $\mathcal{R} = \{\mathcal{R}_k\}_{k=1}^m$, $\cup_{k=1}^m \mathcal{R}_k = \mathcal{E}$, $\mathcal{R}_k \cap \mathcal{R}_j = \varnothing$ $(\forall)k, j = 1 \div m, j \neq k$ ”.

Regula Bayes pentru distribuţii cunoscute. Se presupun cunoscute probabilităţile à priori $\{q_k\}_{k=1}^m$ şi distribuţiile de probabilitate $\{p_k\}_{k=1}^m$, $Y = \{k\}_{k=1}^m$ este mulţimea etichetelor claselor şi $p_Y(\ell) = \Sigma_{k=1}^m q_k \delta_k(\ell)$ este distribuţia de probabilitate pe Y, unde $\delta_k(\ell)$ este funcţia Dirac ($\delta_k(\ell) = 1$ dacă $\ell = k$ şi $\delta_k(\ell) = 0$ în rest). Se numeşte plasator o funcţie $c : \mathcal{E} \to Y$ ce estimează clasa lui $\boldsymbol{x}$, $c(\boldsymbol{x}) = \ell \in Y$ după ce $\boldsymbol{x} \in \mathcal{E}$ a fost observat. Probabilitatea de misclasare pentru clasa k este: $pmc(k) = p[\{c(\boldsymbol{x}) \neq k \mid \{\boldsymbol{x} \in \Pi_k\}\}]$. Funcţia de pierdere discretă pentru plasatorul c faţă de clasa k este: $fpd(c(\boldsymbol{x}), k)$. Riscul funcţional al plasatorului c este

$$rf(c) \equiv E_\mu[fpd(c(\boldsymbol{x}), k)] = \sum_{j=1}^m q_j\, pmc(j) = \sum_{j=1}^m q_j \sum_{k=1, k \neq j}^m \int_{Rj} p_k(\boldsymbol{x})d\boldsymbol{x}$$

deoarece, distribuţia de probabilitate pe $\mathcal{E} \times Y$ este, din construcţie, $\mu(\boldsymbol{x}, k) = q_k\, p_{\ell(x)}(\boldsymbol{x})$ unde cu $\ell(\boldsymbol{x}) \in Y$ s-a notat clasa lui $\boldsymbol{x}$. Dacă se consideră costurile misclasării $\{C(j|k)\}_{k,j=1}^m$ egale cu 1 atunci un plasator va fi optim dacă minimizează $rf(c) = \sum_{k=1}^m q_k \sum_{j=1 j \neq k}^m C(j|k)p(j|k, \mathcal{R})$ adică exact funcţionala din enunţul problemei de clasare. Dacă $\mathcal{X} = \boldsymbol{x}$ probabilitatea à posteriori a clasei k este $p(k|\boldsymbol{x}) = q_k p_k(\boldsymbol{x}) / \sum_{k=1}^n q_k p_k(\boldsymbol{x})$).

Partiţia lui $\mathscr{E}$ care minimizează riscul funcţional $rf(c)$ este $\mathcal{R} = \{\mathcal{R}_k\}^m_{k=1}$, unde regiunile de decizie $\mathcal{R}_k = \{ \boldsymbol{x} \in \mathscr{E} \mid \sum^m_{j=1 j\neq k} q_j\, p_j(\boldsymbol{x}) \leq \sum^m_{j=1 j\neq \ell} q_j\, p_j(\boldsymbol{x}),\ (\forall)\ell\in[1,\ m],\ \ell\neq k \}$ sunt numite regiuni de decizie Bayes, și se înscriu în cazul **3** de separabilitate.

Dacă $p(j|\boldsymbol{x}) = max_{1\leq k\leq m}\ p(k|\boldsymbol{x})$ atunci plasatorul care minimizează riscul funcţional este notat cu $c_B(\boldsymbol{x})$. Plasatorul $c_B(\boldsymbol{x})$ se numeşte plasator Bayes, riscul funcţional pe care acesta îl minimizează se numeşte risc (sau eroare) Bayes, iar partiţia $\mathcal{R}$ care determină şi este determinată de plasatorul Bayes, se numeşte procedură de discriminare (sau clasare) bayesiană. Rezultatul fundamental al analizei discriminante probabiliste clasice este: „dacă $p((p_j(\boldsymbol{x}) / p_\ell(\boldsymbol{x})) = b \mid \boldsymbol{x} \in \Pi_k) = 0,\ (\forall)j,\ k,\ \ell = 1 \div m,\ \ell \neq j$ şi $0 \leq b < \infty$, atunci clasa procedurilor bayesiene este minimală şi completă”.

Regula Bayes pentru distribuții cunoscute permite deci să se construiască o procedură de clasare cu proprietăţi de optimalitate dar aplicabilitatea practică directă este însă redusă deoarece, în realitate, cel puţin distribuţiile $\{p_k\}^m_{k=1}$ nu se cunosc.

Regula de decizie Bayes cu parametrii cunoscuți. Se consideră $m = 2$, cazul a două populații normale, multidimensionale, $\{\Pi_k\}^2_{k=1}$, caracterizate de densităţile de probabilitate: $p_k(\boldsymbol{x}) = (1 / ((2\pi)^{p/2}\, |\boldsymbol{V}|^{1/2}))\ exp[(-1/2)(\boldsymbol{x} - \boldsymbol{\mu}_k)'\, \boldsymbol{V}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_k)]$, adică $[X\in\Pi_k]\Rightarrow[X \sim N(\boldsymbol{\mu}_k,\ \boldsymbol{V})]$, unde $\boldsymbol{\mu}_k \in \mathcal{M}_{p\times 1}(\mathscr{R})$ este vectorul medie şi $\boldsymbol{V} \in \mathcal{M}_{p\times p}(\mathscr{R})$ este matricea de varianţă-covarianţă.

Regiunea de clasificare în Π_1, şi anume $\mathcal{R}_1$, este mulţimea punctelor $\boldsymbol{x} \in \mathscr{R}^p$ pentru care raportul densităţilor $p_1(\boldsymbol{x}) / p_2(\boldsymbol{x}) \geq c$, cu c o constantă convenabil aleasă. Condiţia de definire a lui $\mathcal{R}_1$ revine la: $\mathcal{F}(\boldsymbol{x}) \equiv \boldsymbol{x}'\boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (-1/2)(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)'\, \boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq ln\ c$.

Dacă $\{\Pi_k\}^2_{k=1}$ sunt populaţii multidimensionale, normal distribuite, de medie $\boldsymbol{\mu}_i$ şi cu matricea $\boldsymbol{V}$, de varianţă-covarianţă, comună atunci cele mai bune regiuni de clasificare sunt date de: $\mathcal{R}_1 := \mathcal{F}(\boldsymbol{x}) \geq ln\ c$ și $\mathcal{R}_2 := \mathcal{F}(\boldsymbol{x}) < ln\ c$.

Dacă probabilităţile à priorice q_1 şi q_2 sunt cunoscute, atunci constanta c este dată de relația $c = q_2C(1|2) / q_1C(2|1)$. Dacă $q_1 = q_2$ și $C(1|2) = C(2|1)$ atunci suprafaţa de separare a celor două regiuni este hiperplanul $\mathcal{H}$: $(\boldsymbol{g}_1 - \boldsymbol{g}_2)'(\boldsymbol{x} - (1/2)(\boldsymbol{g}_1 + \boldsymbol{g}_2)) = 0$, unde $\boldsymbol{g}_k = \boldsymbol{V}^{-1}\boldsymbol{\mu}_k$ este prototipul populaţiei Π_k iar clasificatorul obţinut este un clasificator cu distanţă minimă.

Dacă probabilităţile à priorice nu sunt cunoscute atunci $C = ln\ c$ va fi ales astfel încât: $C(1|2)(1-\Phi((C + (1/2)\alpha) / \sqrt{\alpha})) = C(2|1)(\Phi((C-(1/2)\alpha) / \sqrt{\alpha}))$, unde C(k|j) sunt cele două costuri ale misclasării, $\alpha = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)'\boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ este distanţa Mahalanobis dintre cele două populaţii iar $\Phi(x) = \int_{-\infty}^{x} (1/\sqrt{2\pi})e^{\varphi(t)}dt$ cu $\varphi(t) = -(t^2/2)$, este funcţia de repartiţie a variabilei aleatoare Gauss-Laplace.

Regula de decizie Bayes cu parametrii necunoscuți. În cazul în care probabilităţile à priori nu sunt cunoscute, se generează o clasă de proceduri admisibile pe bază de estimații.

Dacă $\boldsymbol{x}_1^{(i)}, ..., \boldsymbol{x}_{n_i}^{(i)} \in N(\boldsymbol{\mu}_i, \boldsymbol{V})$, $i \in \{1, 2\}$ sunt două selecţii bernoulliene atunci estimatorii $\bar{\boldsymbol{x}}_i = (1/n_i) \sum^{n_i}_{j=1} \boldsymbol{x}^{(i)}_j$ şi $((n_1 -1) + (n_2 -1))\boldsymbol{S} = (n_1 + n_2 - 2)\boldsymbol{S} = \sum^2_{i=1} \sum^{n_i}_{j=1}(\boldsymbol{x}^{(i)}_j - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}^{(i)}_j - \bar{\boldsymbol{x}}_i)'$ sunt estimatori nedeplasaţi, de verosimilitate maximă, ai lui $\boldsymbol{\mu}_i$, şi $\boldsymbol{V}$. Pentru selecţii suficient de mari folosirea estimaţiilor în locul valorilor exacte implică erori mici.

Substituind parametrii estimaţi în relaţiile de definiţie ale regiunilor de decizie se obţine: $\mathcal{R}_1 := \tilde{\mathcal{F}}(\boldsymbol{x}) \geq ln\ c$ și $\mathcal{R}_2 := \tilde{\mathcal{F}}(\boldsymbol{x}) < ln\ c$, unde $\tilde{\mathcal{F}}(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) - (1/2)(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})$.

Dacă se dorește clasificarea selecţiilor reunite ca un tot, atunci se utilizează următorii estimatori, respectiv criteriu: $n = n_1 + n_2$, $\bar{\boldsymbol{x}} = (1/n)\sum^n_{j=1}\boldsymbol{x}_j$ cu $[\ \boldsymbol{x}_j \in \Pi_1\ ] \vee [\ \boldsymbol{x}_j \in \Pi_2\ ]$ și $(n_1 + n_2 + n - 3)\bar{\boldsymbol{S}} = \boldsymbol{S} + \sum^n_{j=1}(\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})'$. $\mathcal{R}_1 := (\bar{\boldsymbol{x}} - (1/2)(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2))'\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \geq c$.

Prospectorul de date poate obține diverse particularizări ale regiunilor de decizie Bayes pentru diverse valori privind numărul m de populații și numărul p de variabile sau pentru diverși estimatori de verosimilitate maximă definiți în cadrul unor ipoteze compozite.

Estimare bayesiană. In abordările anterioare (frecventiste) s-a presupus o selecţie aleatoare dintr-o populaţie având densitatea de probabilitate $f(\boldsymbol{x}; \theta)$ cu $\boldsymbol{x} \in \boldsymbol{X}$ şi $\theta \in \Theta$. O procedură de inferenţă frecventistă depinde de funcţia de verosimilitate $L(\theta) = \prod^{n}_{i=1} f(x_i; \theta)$, unde θ este necunoscut dar fixat.

In demersul bayesian se presupune à priori că parametrul necunoscut θ este o variabilă aleatoare având o distribuţie de probabilitate proprie pe spaţiul Θ al parametrilor, notată $h(\theta)$ şi numită distribuţia à priorică a lui θ, $f(x; \theta)$ devenind $f(x|\theta)$. Distribuţia à priorică este, în cazul ideal, fixată înainte de începerea culegerii selecţiei bernoulliene.

Dacă $f(\boldsymbol{x}|\theta)h(\theta)$, distribuţia comună a lui $\boldsymbol{x}$ şi θ, şi $m(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\theta)h(\theta)d\theta$, distribuţia marginală a lui $\boldsymbol{x}$, sunt cunoscute, atunci distribuţia lui θ condiţionată de evenimentul $X=\boldsymbol{x}$ sau distribuţia à posteriori a lui θ, este: $h(\theta|\boldsymbol{x})=h(\theta|X=\boldsymbol{x}) = f(\boldsymbol{x}|\theta)h(\theta) / m(\boldsymbol{x})$, $m(\boldsymbol{x})>0$, $\boldsymbol{x}\in\mathscr{E}$, $\theta\in\Theta$.

Dacă $\theta \sim N(\boldsymbol{m}, \boldsymbol{S})$ şi $\boldsymbol{x} \sim N(\theta, \boldsymbol{V})$, atunci $h(\theta|\boldsymbol{x})$ este densitatea de probabilitate a unei $N(\mu, \boldsymbol{C})$ cu $\mu = \boldsymbol{S}(\boldsymbol{S} + \boldsymbol{V})^{-1}\boldsymbol{x} + \boldsymbol{V}(\boldsymbol{S} + \boldsymbol{V})^{-1}\boldsymbol{m}$ şi $\boldsymbol{C} = \boldsymbol{V}(\boldsymbol{S} + \boldsymbol{V})^{-1}\boldsymbol{S}$.

Dacă $\theta \sim N(\tau, \sigma^2_0)$ şi $x \sim N(\theta, \sigma^2_1)$, atunci densitatea à posteriori a lui θ este: $N(\mu, \sigma^2)$, unde $\mu = (x/\sigma^2_1 + \tau/\sigma^2_0)\,(1/\sigma^2_0 + 1/\sigma^2_1)^{-1}$ şi $\sigma^2 = (\sigma^2_0\,\sigma^2_1)/(\sigma^2_0 + \sigma^2_1) = (1/\sigma^2_0 + 1/\sigma^2_1)^{-1}$.

Pentru variabila aleatoare X, cu densitatea de probabilitate $f(x, \theta)$, funcţia $T : \Omega \to \mathscr{R}$ se numeşte statistică suficientă pentru $\theta \Leftrightarrow f(\boldsymbol{x}|T(\boldsymbol{x}) = t, \theta) = f(\boldsymbol{x}|T(\boldsymbol{x}) = t)$ $(\forall)t \in \Delta \subseteq \mathscr{R}$, adică dacă și numai dacă densitatea de probabilitate condiţionată a lui X este independentă de θ.

Fie $\boldsymbol{X} = (\boldsymbol{x_1}, ..., \boldsymbol{x_n})$ o selecţie bernoulliană asupra unei variabile aleatoare ce depinde de θ și fie $\delta \equiv \delta(T)$ un estimator a lui θ. Funcţia de pierdere, ce se obţine estimând θ prin δ, este: $L^{\boldsymbol{b}}(\theta, \delta) \equiv L^{\boldsymbol{b}}(\theta, \delta(T)) = (\delta(T) - \theta)^2$. $R^{\boldsymbol{b}}(\theta, \delta) = E[L^{\boldsymbol{b}}(\theta, \delta)] = \int_{\Delta} L^{\boldsymbol{b}}(\theta, \delta(t))\, f(t|\theta)dt$, este riscul funcțional. Se numeşte risc bayesian: $r^{\boldsymbol{b}}(\theta, \delta) = \int_{\Theta} R^{\boldsymbol{b}}(\theta, \delta)h(\theta)d\theta$. Se numeşte estimator bayesian $r^{\boldsymbol{b}}(\theta, \delta^{\boldsymbol{b}}) = \inf_{\delta\in\boldsymbol{B}} r^{\boldsymbol{b}}(\theta, \delta)$, $\delta^{\boldsymbol{b}}\in B$, unde B este clasa estimatorilor pentru care riscul bayesian este finit. In cazul funcţiei de pierdere „suma pătratelor erorilor” estimatorul bayesian este $\delta^{\boldsymbol{b}}(t) = \int_{\Theta} \theta h(\theta|t)d\theta \equiv E[\theta|T(\boldsymbol{x}) = t]$, adică media distribuţiei à posteriori $h(\theta|t)$ pentru toate valorile posibile observate $t \in \Delta$.

Fie $\boldsymbol{x_1}, ..., \boldsymbol{x_n}$ variabile aleatoare independente şi identic repartizate $N(\theta, \sigma^2_1)$ cu θ necunoscut şi $\sigma_1 > 0$ dat și fie statistica $T = (1/n)\Sigma^{n}_{i=1} \boldsymbol{x_i}$, care este suficientă pentru θ. Dacă distribuţia a priori a lui θ pe spaţiul $\Theta = \mathscr{R}$ este $N(\tau, \sigma^2_0)$ cu $\tau, \sigma_0 \in \mathscr{R}$ daţi şi $\sigma_0 > 0$, atunci distribuţia à posteriori a lui θ, condiţionată de observaţiile $\boldsymbol{x_1}, ..., \boldsymbol{x_n}$, este $N(\mu, \sigma^2)$, unde $\mu = (\,(n\sigma^2_0\,)/(n\sigma^2_0 + \sigma^2_1)\,)T(\boldsymbol{x}) + ((\sigma^2_1\,)/(n\sigma^2_0 + \sigma^2_1))\tau$ şi $\sigma^2 = (\sigma^2_0\,\sigma^2_1)/(n\sigma^2_0 + \sigma^2_1)$. Dacă $\sigma_0 = 0$, atunci $\mu = \tau$ indiferent de observaţiile efectuate. Dacă $\sigma_0 > \sigma_1$ rezultă $\mu \approx \bar{x}$, cunoaşterea mediei a priorice τ este de importanţă redusă. Raportul $a = \sigma^2_1 / \sigma^2_0$ măsoară încrederea à priori că τ este o estimare corectă a mediei. Dacă $a < \infty$ atunci $\lim_{n\to\infty} \mu = \lim_{n\to\infty} \bar{x}$. Dacă dispersia iniţială este mică, media estimată tinde să rămână în apropierea mediei iniţiale τ chiar dacă media empirică $\bar{x}$ diferă considerabil de aceasta. Dacă raportul a este mic, atunci media şi dispersia à priori au doar o influenţă redusă asupra estimării parametrilor care sunt determinaţi aproape exclusiv din datele empirice. Dacă $T(\boldsymbol{x}) = t$, estimatorul Bayes al mediei unei variabile aleatoare $N(\mu, \sigma^2)$ este: $\delta(t) = \tilde{\theta}_B = (nt/\sigma^2_1 + nt/\sigma^2_0)\,(1/\sigma^2_1 + 1/\sigma^2_0)^{-1}$. Pentru cazul multidimensional se obţine: $\tilde{\theta}_B = \boldsymbol{S}(\boldsymbol{S} + (1/n)\boldsymbol{V})^{-1}\boldsymbol{t} + (1/n)\boldsymbol{V}(\boldsymbol{S} + (1/n)\boldsymbol{V})^{-1}\boldsymbol{m}$.

Fie $\boldsymbol{X} = (\boldsymbol{x_1}, ..., \boldsymbol{x_n})$ o selecţie bernoulliană din populaţiile Π_1 şi Π_2. Dacă $X \in \Pi_1$, atunci densitatea de probabilitate este $f_i(\boldsymbol{x}|\theta)$, $\theta \in \theta_i$ şi densitatea à priorică este $h_k(\theta)$, $k = 1 \div 2$. Dacă q_1 și q_2 sunt probabilităţile à priori ale populaţiilor Π_1, și Π_2, probabilităţile à posteriori

sunt: $p(\Pi_k \mid \boldsymbol{x}) = m_k(\boldsymbol{x})q_k \;/\; (m_k(\boldsymbol{x})q_k + m_k(\boldsymbol{x})q_k)$, unde $m_k(\boldsymbol{x})$ este densitatea de probabilitate marginală a lui $\boldsymbol{x}$ condiționat de faptul că provine din Π_k: $m_k(\boldsymbol{x}) = \int_{\Theta k} f_k(x|\theta)\, h_k(\theta)\, d\theta$, $k = 1 \div 2$.

Procedura bayesiană de discriminare este:

$$\boldsymbol{x} \in \begin{cases} \Pi_1 & \text{dacă } p(\Pi_1 \mid \boldsymbol{x}) / p(\Pi_2 \mid \boldsymbol{x}) = (q_1/q_2)B_{12}(\boldsymbol{x}) \geq 1 \\ \Pi_2 & \text{în caz contrar} \end{cases}$$

unde $B_{12}(\boldsymbol{x}) = m_1(\boldsymbol{x}) / m_2(\boldsymbol{x})$ este cunoscut ca factorul Bayes al populației Π_1 versus Π_2.

Mașini cu suport vectorial

Mașinile cu suport vectorial reprezintă o clasă de algoritmi de învățare destinați, inițial, problemelor de discriminare adică de predicție unei variabile calitative. Ulterior, algoritmii au fost generalizați pentru a prezice o variabilă cantitativă adică de a găsi o funcție de discriminare (sau clasificator) a cărei capacitate de generalizare (sau calitate a predicției) să fie cea mai mare posibilă. Abordarea s-a concentrat pe proprietățile de generalizare (sau de previziune) ale unui model controlându-i complexitatea, mai precis, integrând în estimare numărul de parametri, în acest caz numărul de vectori suport. Ideea de bază al mașinilor cu suport vectorial a fost de a reduce problema discriminării la o problemă, liniară, de căutare a unui hiperplan optimal: fie, prin definirea hiperplanului optimal ca soluție a unei probleme de optimizare cu restricții, în care funcția obiectiv se exprimă numai cu ajutorul produselor scalare între vectori iar numărul de restricții „active” (vectorii suport) controlează complexitatea modelului, fie prin căutarea unor suprafețe de separare neliniare, fie prin introducerea unei funcții nucleu în produsul scalar inducând implicit o transformare neliniară a datelor către un spațiu Hilbert, intermediar, de dimensiune mai mare și în care este rezolvată problema liniară.

Fie Y variabila de explicat și fie $X = \{X^j\}_{j=1}^{p}$ variabilele explicative sau de predicție. X este o variabilă cu valori într-o mulțime $\mathscr{E} \subset \mathscr{R}^p$ iar $\varphi(\boldsymbol{x})$ este un model pentru Y , adică o funcție $\varphi : \mathscr{E} \to \mathscr{B}$, unde $\boldsymbol{x} = (x_j)_{j=1}^{p} \in \mathscr{E}$ și $\varphi(\boldsymbol{x}) \in \mathscr{B} \subset \mathscr{R}$.

Se presupune că: variabila Y este dicotomică, $\mathscr{B} = \{-1, 1\}$ și $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ este un eșantion statistic de mărime n și de lege F necunoscută. Obiectivul este de a construi o estimare $\tilde{\varphi} : \mathscr{E} \to \{-1, 1\}$ a lui φ astfel încât probabilitatea $p(\varphi(X) \neq Y)$ să fie minimă. Problema revine la a căuta o frontieră de decizie în spațiul $\mathscr{E}$ pentru valorile lui X și la a găsi un compromis între complexitatea acestei frontiere, respectiv, capacitatea de ajustare a modelului, și calitățile de generalizare (sau de previziune) ale modelului. Demersul constă în a găsi o funcție reală f al cărui semn să ofere previziunea: $\tilde{\varphi} = sign(f)$. Eroarea de previziune se exprimă prin cantitatea: $p(\varphi(X) \neq Y) = p(Yf(X) \leq 0)$. Valoarea absolută a acestei cantități, $|Yf(X)|$, furnizează o indicație privind încrederea care poate fi acordată rezultatului clasării. Se spune că $Yf(X)$ este marja lui f în (X, Y). Primul pas este de a transforma valorile lui X, adică obiectele din $\mathscr{E}$, printr-o funcție $\Phi : \mathscr{E} \to \mathscr{H}$ cu valori într-un spațiu $\mathscr{H}$, intermediar, înzestrat cu un produs scalar. Această transformare, fundamentală pentru abordarea *SVM*, ia în considerare eventuala neliniaritate a problemei de rezolvat și conduce la rezolvarea unei separări liniare.

În cazul în care Φ este funcția identitate (adică în cazul liniar), atunci când separarea este posibilă, dintre toate hiperplanele, soluții de separare a observațiilor, se alege acela care este situat „cel mai departe” de toate exemplele, adică de marjă maximală. Cu produsul scalar al spațiului $\mathscr{H}$, un hiperplan $\mathcal{H}$ este definit prin ecuația $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$, unde $\boldsymbol{w}$ este un vector ortogonal pe hiperplan, $\boldsymbol{w} \perp \mathcal{H}$, iar semnul funcției $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$ arată de care parte a

hiperplanului este situat punctul $\boldsymbol{x}$ de explicat. Un punct $\boldsymbol{x}$ este bine clasat $\Leftrightarrow yf(\boldsymbol{x}) \geq 1$. Un hiperplan $\mathcal{H} \equiv (\boldsymbol{w}, b)$ este un separator dacă: $y_i f(\boldsymbol{x}_i) \geq 1\ (\forall) i \in [1, n]$. Distanța de la un punct $\boldsymbol{x}$ la $(\boldsymbol{w}, b)$ este: $d(\boldsymbol{x}) = |\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b| / \|\boldsymbol{w}\| = |f(\boldsymbol{x})| / \|\boldsymbol{w}\|$ iar marja hiperplanului are valoarea $2 / \|\boldsymbol{w}\|^2$. Căutarea hiperplanului separator de marjă maximală revine la rezolvarea problemei (primare) de optimizare cu restricții: $(1/2) min_{\boldsymbol{w}} \|\boldsymbol{w}\|^2 \ ¦\ y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1\ (\forall) i$. Problema duală se obține prin introducerea multiplicatorilor Lagrange. Soluția este furnizată de un punct șa $(\boldsymbol{w}^*, b^*, \lambda^*)$ al lagranjianului $L(\boldsymbol{w}, b, \lambda) = (1/2) \| \boldsymbol{w} \|^2 - \sum^{n}_{i=1} \lambda_i (y_i (\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1)$, punctul șa verificând condițiile: $\lambda^*_i [y_i (\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b^*) - 1] = 0\ \ (\forall) i \in [1, n]$. Vectorii suport sunt vectorii $\boldsymbol{x}_i$ pentru care restricția este activă (cele mai aproapiate de hiperplan) adică verifică: $y_i(\langle \boldsymbol{w}^*, \boldsymbol{x}_i \rangle + b^*) = 1$. Condițiile de anulare a derivatelor parțiale permit exprimarea formulei duale a lagranjianului: $W(\boldsymbol{\lambda}) = \sum^{n}_{i=1} \lambda_i - (1/2) \sum^{n}_{i=1} \sum^{n}_{j=1} \lambda_i \lambda_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$. Pentru a găsi punctul de șa, se maximizează $W(\boldsymbol{\lambda})$, $\lambda_i \geq 0\ (\forall) i \in [1, n]$. Rezolvarea acestei probleme de optimizare pătratică de dimensiune n (numărul de observații), furnizează ecuația hiperplanului optimal: $\sum^{n}_{i=1} \lambda^*_i y_i = \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + b^* = 0$ cu $b_0 = -(1/2)(\langle \boldsymbol{w}^*, \boldsymbol{SV}_{clasa+1} \rangle + \langle \boldsymbol{w}^*, \boldsymbol{SV}_{clasa-1} \rangle)$. Pentru o nouă observație $\boldsymbol{x}$ prezentată modelului, este suficient să se vadă semnul expresiei $f(\boldsymbol{x}) = \sum^{n}_{i=1} \lambda^*_i y_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + b^*$ pentru a ști în care semi-spațiu se află $\boldsymbol{x}$ și deci ce clasă i se va atribui.

Dacă observațiile nu sunt separabile printr-un hiperplan atunci se recurge la o „relaxare” a restricțiilor introducându-se termenii de eroare, ξ_i, $y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \geq 1 - \xi_i\ \ (\forall) i \in [1, n]$, care controlează depășirile. Modelul va oferi un răspuns greșit pentru un vector $\boldsymbol{x}_i$ dacă valoarea termenului de eroare corespunzător este mai mare decât 1, $\xi_i > 1$. Introducând o penalizare δ pentru încălcarea restricțiilor, problema de minimizare se reformulează în felul următor $min (1/2) \|\boldsymbol{w}\|^2 + \delta \sum^{n}_{i=1} \xi_i \ ¦\ y_i \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b \geq 1 - \xi_i$, $(\forall) i \in [1, n]$. Problema se formulează în aceeași formă duală ca și în cazul separabilității cu o singură diferență: coeficienții λ_i sunt mărginiți de constanta δ de control a penalizării. Din punctul de vedere al prospectorului de date parametrul δ, care controlează penalizarea, trebuie „bine” ales fiind parametrul care reprezintă compromisul între o bună ajustare și o bună generalizare. Cu cât el este mai mare cu atât importanța atribuită ajustării modelului este mai puternică.

Observațiile făcute în mulțimea $\mathcal{E}$ (de obicei, $\mathcal{R}^p$) sunt transformate prin aplicația neliniară $\Phi : \mathcal{E} \rightarrow \mathcal{H}$, spațiul $\mathcal{H}$ fiind de dimensiune mai mare și înzestrat cu un produs scalar. Formularea problemei de minimizare și soluția sa: $f(\boldsymbol{x}) = \sum^{n}_{i=1} \lambda^*_i y_i \langle \boldsymbol{x}, \boldsymbol{x}_i \rangle + b^*$ implică numai elementele $\boldsymbol{x}$ și $\boldsymbol{x}'$, prin intermediul produsului scalar $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$. Prin urmare, nu ar mai fi necesară explicitarea transformării Φ, ceea ce de multe ori este imposibil, cu condiția de a dispune de o exprimare a produselor scalare în $\mathcal{H}$ cu ajutorul unei funcții $k : \mathcal{E} \times \mathcal{E} \rightarrow \mathcal{R}$, simetrică, numită nucleu (*kernel*), astfel încât: $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x}') \rangle$. Convenabil ales, nucleul permite materializarea unei noțiuni de „proximitate”, adaptată problemei de discriminare și structurii sale de date. Pentru construirea de funcții nucleu se recurge la combinări ale unor nuclee simple: fie liniare $k(\boldsymbol{x}', \boldsymbol{x}'') = \langle \boldsymbol{x}', \boldsymbol{x}'' \rangle$, fie polinomiale $k(\boldsymbol{x}', \boldsymbol{x}'') = (c + \langle \boldsymbol{x}', \boldsymbol{x}'' \rangle)^d$ sau gaussiene $k(\boldsymbol{x}', \boldsymbol{x}'') = e^{-\alpha(\boldsymbol{x}', \boldsymbol{x}'')}$, unde $\alpha(\boldsymbol{x}', \boldsymbol{x}'') = \| \boldsymbol{x}' - \boldsymbol{x}'' \|^2 / 2\sigma^2$, pentru a se obține nuclee mai complexe (multidimensionale) asociate cu situația întâlnită. Pentru prospectorul de date, o mare flexibilitate în definirea nucleelor, care să permită definirea unor noțiuni adecvate de similitudine, conferă mai multă eficacitate acestei abordări cu condiția, desigur, de a construi și a testa un nucleu „bun”. Rezultă, din nou, importanța unei evaluări corecte a erorilor de previziune, de exemplu, prin validare încrucișată.

Metode conexioniste.

O rețea neuronală este asocierea într-un graf, mai mult sau mai puțin complex, a neuronilor formali. Neuronul formal este un model al neuronului biologic care se caracterizează prin: stări interne, $s \in S$, semnale de intrare, $\{x_i\}^p_{i=1}$, funcția de tranziție a stărilor $s = h(x_1, \ldots, x_p) = f(\beta_0 + \sum^p_{j=1}\beta_j x_j)$. Valorile coeficienților $\{\beta_j\}^p_{j=0}$ sunt estimate într-o fază de învățare și constituie „memoria" sau „cunoașterea distribuită" a rețelei, coeficientul β_0 este numit *bias* al neuronului. Rețelele neuronale sunt caracterizate prin organizarea grafului (în straturi), prin numărul de neuroni și prin tipul neuronilor, respectiv, funcțiile lor de tranziție. Perceptronul multistrat este o rețea formată din straturi succesive de neuroni formali; stratul este un set de neuroni fără nici-o legătură între ei; stratul de intrare citește semnalele $\{x_j\}^p_{j=1}$ de intrare și conține câte un neuron pentru fiecare intrare x_j; unul sau mai multe straturi ascunse participă la transfer, un neuron al unui strat ascuns este conectat la intrare cu fiecare dintre neuronii stratului precedent și la ieșire cu fiecare neuron al stratului următor; stratul de ieșire furnizează răspunsul sistemului. Un perceptron multistrat realizează o transformare $y = \varphi(x_1, \ldots, x_p; \boldsymbol{\beta})$ unde $\boldsymbol{\beta}$ este vectorul conținând parametrii $\beta_{jk\ell}$ corespunzători intrării j a neuronului k din stratul ℓ; stratul de intrare ($\ell = 0$) nu este parametrizat pentru că nu face altceva decât să distribuie intrările în neuronii din stratul următor. Intrările rețelei $\{x_i\}^p_{i=1}$, sunt variabilele explicative ale modelului, ieșirea y este variabila de explicat (dependentă sau țintă) iar $\boldsymbol{\beta}$, vectorul ponderilor intrărilor în fiecare neuron al rețelei, reprezintă parametrii de estimat în urma unui proces de învățare.

Pentru un eșantion de învățare $\{(x^1_i, \ldots, x^p_i ; y_i)\}^n_{i=1}$ construit din n observații asupra a p variabile explicative $\{X^j\}^p_{j=1}$ și a unei variabile de explicat Y, învățarea constă în estimarea vectorului de parametri $\boldsymbol{\beta}$ rezolvând o problemă a celor mai mici pătrate: $\tilde{\boldsymbol{\beta}} = min_{\boldsymbol{b}}\, Q(\boldsymbol{b})$, unde: $Q(\boldsymbol{b}) = (1/n)\sum^n_{i=1}(y_i - \varphi(x^1_i, \ldots, x^p_i; (\boldsymbol{b})))^2$. Algoritmul de optimizare cel mai utilizat este un algoritm de retropropagare (propagare inversă) a gradientului bazat pe faptul că în orice punct $\boldsymbol{b}$ vectorul gradient al lui Q este orientat în direcția de creștere a erorii și deci pentru a-l descrește pe Q este suficientă o deplasare în sens contrar. Pornind de la erorile observate pe ieșiri, formula retropropagării erorii furnizează expresia erorii atribuite fiecărei intrări, de la stratul de ieșire către stratul de intrare. Proprietățile acestui algoritm implică o convergență aproape sigură, probabilitatea de atingere a unei precizii dorite (fixate à priori) tinde către *1* atunci când dimensiunea eșantionului de învățare tinde către infinit.

În practică, prospectorul de date se confruntă cu o serie de opțiuni privind, în principal, controlul supra-învățării: alegerea unor parametri (limitarea numărului de neuroni, limitarea duratei de învățare, creșterea coeficientului de penalizare a normei parametrilor); alegerea modului de estimare a erorii (pe eșantionul de test sau validare încrucișată).

Metoda segmentării

Metoda segmentării este o metodă complementară de rezolvare a problemelor de discriminare și de regresie prin împărţirea progresivă a eşantionului de observaţii într-un arbore de decizie binară.

Fie $\boldsymbol{y}$ variabila privilegiată, discretă, cu q modalităţi, $\{k\}^q_{k=1}$, care este explicată prin variabilele, cantitative sau calitative, $\{X^j\}^p_{j=1}$, și fie $\{\boldsymbol{x_i}\,; y_i)\}^n_{i=1} \equiv \{\{x^j_i\}^p_{j=1}; y_i)\}^n_{i=1}$ eșantionul observațiilor, unde $y_i \in \{k\}^q_{k=1}$. Metoda de segmentare constă, mai întâi, în a căuta variabila X^j care, explică cel mai bine variabila $\boldsymbol{y}$ și defineşte o împărţire a eşantionului în două submulţimi de indivizi, numite segmente sau noduri. Apoi, se reiterează procedeul căutându-se cea mai bună variabilă în interiorul fiecăruia dintre cele două segmente definite, ş.a.m.d.

Prin împărţirea succesivă a eşantionului în câte două submulţimi rezultă un arbore de decizie binară în care se disting: segmente intermediare, segmente terminale, ramuri ale unui segment, arborele binar complet, A_{max}, și subarbori. Efectuarea diviziunii unui nod se face astfel încât cele două segmente descendente să fie mai omogene decât nodul părinte şi cât mai diferite între ele faţă de variabilă. Fazele de construire ale arborelui de decizie binară sunt: stabilirea, pentru fiecare nod, a mulţimii diviziunilor admisibile; definirea unui criteriu de selecţionare a „celei mai bune" diviziuni a fiecărui nod; definirea unei reguli care să permită declararea unui nod ca terminal sau intermediar; afectarea fiecărui nod terminal unei clase; estimarea riscului de misclasare.

Iniţial, există un singur segment conţinând toţi indivizii $\boldsymbol{x_i}$, $i = 1 \div n$. Sunt examinate, secvenţial, toate variabilele explicative X^j , $j = 1 \div p$. În funcție de natura fiecărei variabile X^j (continuă sau discretă) se definesc toate diviziunile posibile. O diviziune posibilă este admisibilă dacă segmentele descendente sunt nevide. Dintre toate diviziunile admisibile ∂^j_m, unde m reprezintă a m-a diviziune (sau a m-a valoare ordonată a variabilei din eşantion), este selecţionată diviziunea $\tilde{\partial}^j$ „cea mai bună" în sensul unui criteriu de impuritate. Astfel, pentru fiecare din cele p variabile, se obține diviziunea optimă „locală" $\tilde{\partial}^j$ şi, în final, din cele p diviziuni se va reține diviziunea $\tilde{\partial}$, care va furniza cele două segmente „cele mai caracteristice" vis-à-vis de $\boldsymbol{y}$. Procedeul se aplică iterativ fiecărui segment descendent obţinut și se opreşte când toate segmentele sunt declarate terminale. Afectarea unui individ nou se face prin „coborârea" lui pe ramurile arborelui.

Fie $p(r|a)$ probabilitatea condiţionată de apartenenţă la grupul G_r, $r \in \{1, 2, ..., q\}$ a mulţimii observaţiilor din nodul a. Impuritatea unui nod, a, este o funcţie nenegativă de $\{p(r|a)\}^q_{r=1}$, care verifică următoarele condiţii: este maximală când probabilităţile de apartenenţă la diferite grupuri sunt egale între ele: $p(r|a) = 1$, $(\forall)r$; este nulă dacă nodul conţine observaţii aparţinând unui singur grup: $p(r|a) = 1$ şi $p(s|a) = 0$, $(\forall)r, s, s \neq r$; este o funcţie simetrică de probabilităţile $p(r|a)$. Funcţiile de impuritate, cele mai des utilizate, sunt: $i(a) = -\sum^q_{r=1} p(r|a)\, ln(p(r|a))$, funcţie derivată din noţiunea de entropie Shannon și indicele de diversitate Gini $i(a) = -\sum_{r \neq s} p(r|a)\, p(s|a)$.

Fie ∂ o diviziune admisibilă care împarte nodul a în segmentele t_s şi t_d cu probabilităţile: $p_s \equiv p(\, t_s|a\,) = p(t_s) / p(a)$ și respectiv $p_d \equiv p(\, t_d|a\,) = p(t_d) / p(a)$. Reducerea impurităţii nodului a datorată diviziunii ∂ este definită prin expresia : $\Delta_i(\partial, a) = i(a) - p_s\, i(t_s) - p_d\, i(t_d)$. Orice diviziune, ∂, a unui nod, a, duce la o reducere pozitivă sau nulă a impurităţii. Cea mai „bună" diviziune este $\tilde{\partial}^j = argmax_{m \in \partial j}\, \Delta_i(\partial^j_m, t)$ adică aceea pentru care reducerea impurităţii este maximă, unde ∂^j este mulţimea diviziunilor admisibile ale variabilei X^j. Pe mulţimea $\{X^j\}^p_{j=1}$, a tuturor variabilelor explicative, diviziunea nodului t este efectuată cu ajutorul variabilei X^j care asigură $\tilde{\partial} = max_{1 \leq j \leq p}\{\tilde{\partial}^j\}$.

În procesul de construire a lui A_{max} este posibil ca toate nodurile terminale, a, ale arborelui curent, A, să fie afectate unuia din cele q grupuri (sau clase). Fiecărei erori de clasare i se asociază un preţ de misclasare $\gamma(s/r)$, $s, r = 1 \div q$, costul misclasării fiind $\sum^q_{r=1} \gamma(s/r) p(r|a)$.

Un nod a va fi asignat acelei clase $\tilde{s}$ pentru care $\tilde{s} = min_{1 \leq s \leq q} \sum^q_{r=1} \gamma(s/r)\, p(r|a)$. Dacă minimul este atins pentru cel puţin două clase atunci nodul este afectat arbitrar uneia dintre aceste clase. Dacă $\gamma(s/r) = 1$, $(\forall)s \neq r$ şi $\gamma(s/s) = 0$, $(\forall)s$, atunci nodul va fi asignat clasei cu cei mai mulţi reprezentanţi în ea. Costul misclasării unei observaţii aparţinând nodului a este: $c(a) = min_{1 \leq s \leq q} \sum^q_{r=1} \gamma(s/r) p(r|a)$. Costul misclasării datorat nodului a, este $C(a) = c(a)\, p(a)$, unde $p(a)$ este probabilitatea nodului. Riscul erorii de afectare datorat arborelui A, $rea(A)$, este: $rea(A) = \sum_{a \in \acute{A}} C(a) = \sum_s \sum_{a \in \acute{A}(s)} \sum_r \gamma(s/r)\, p(r|a) p_r = \sum_s \sum_r \gamma(s/r)(n_{sr}/n)$, unde $\acute{A}$ este mulţimea

nodurilor terminale ale lui A, $\acute{A}(s)$ este mulţimea nodurilor terminale ale lui A asignate clasei s, p_r este probabilitatea à priori ca un nod să provină din clasa r, n_{sr} este numărul de indivizi din clasa r clasaţi în clasa s, $s \neq r$.

Un subarbore al lui A_{max} este optimal („cel mai bun") dacă numărul de segmente terminale conținute și riscul erorii de afectare sunt minime şi, în plus, furnizează o estimaţie corectă a erorii teoretice de clasare. Pentru selecţia subarborelui optimal se împarte eşantionul iniţial într-un eşantion de învăţare şi un eşantion de testare. Pornind de la eşantionul de învăţare se construieşte arborele A_{max}. Operaţia de „tundere" a arborelui A_{max} constă în construirea unui şir optimal A_H, ..., A_h, ..., A_1 de subarbori incluşi, unde A_H este A_{max}, A_h este subarborele cu h segmente terminale, A_1 este eşantionul total. Fiecare subarbore A_h din acest şir este optimal în sensul că eroarea aparentă a subarborelui este minimală printre toţi subarborii având acelaşi număr de segmente terminale, adică $ea(A_h) = min_{A \in S_h} ea(A)$, unde S_h este mulţimea subarborilor lui A_{max} cu h segmente terminale. Se selectează din şirul de arbori optimali subarborele $\tilde{A}$ cu eroarea teoretică minimă, adică $et(\tilde{A}) = min_{1 \le h \le H} et(A_h)$. Eroarea teoretică se estimează după formula $\tilde{et}(A) = \sum_{t \in A} \tilde{R}_t$, unde: $\tilde{R}_t = (\tilde{n}_t / \tilde{n}) \times \tilde{s}^2_t$, $\tilde{n}$ este volumul eşantionului de test, $\tilde{n}_t$ este numărul de indivizi din eşantionul de test aparţinând segmentului t, $\tilde{\bar{y}}$ este media de selecţie în interiorul segmentului t și $\tilde{s}^2_t = (1 / \tilde{n}_t) \sum^{|t|}_{i=1} (y_i - \tilde{\bar{y}})$ este dispersia de selecţie a variabilei y în interiorul segmentului t, $|t| = card(t)$.

Deși cea mai bună diviziune, $\tilde{\delta}$, a unui nod este cea care asigură cea mai mare reducere a dispersiei reziduale (sau a impurității), prin trecerea de la acel nod la segmentele descendente, prospectorul de date poate utiliza și alte diviziuni (echi-reductive, echi-divizante), aproximativ la fel de bune, dar foarte importante la nivelul interpretării.

Metode de agregare a modelelor

Agregarea (sau combinarea) unui număr mare de modele permite ameliorarea ajustării modelelor definite prin arbori decizionali evitându-se, totodată, supraajustarea acestora și se bazează pe două tipuri de strategii de agregare: aleatoare (*bagging*) și adaptive (*boosting*).

Strategii aleatoare. Principiul *bagging*-ului se bazează pe faptul că medierea previziunilor mai multor modele independente permite reducerea varianţei şi deci reducerea erorii de previziune.

Fie Y variabila de explicat, cantitativă sau calitativă cu modalităţile $\tau = 1 \div q$, fie $X = \{X^j\}^p_{j=1}$ variabilele explicative, fie $\varphi(X)$ un model funcţie de X și fie $\boldsymbol{z} = \{(x_i, y_i)\}^n_{i=1}$ un eşantion de lege F. Speranţa, $E_F(\tilde{\varphi}_z)$, a unui estimator $\tilde{\varphi}_z$ definit pe eşantionul $\boldsymbol{z}$, este un estimator fără *bias*, de varianţă nulă.

Se consideră K eşantioane independente, notate $\{\boldsymbol{z}_\kappa\}^K_{\kappa=1}$, şi se construieşte familia de modele $\{\varphi_{z_\kappa}\}^K_{\kappa=1}$. Estimarea medie va fi:

$$\tilde{\varphi}_K(\bullet) = \begin{cases} E_F(\tilde{\varphi}_{z_\kappa}) = (1 / K) \sum^K_{\kappa=1} \tilde{\varphi}_{z_k}(\bullet), \text{ dacă variabila de explicat } Y \text{ este cantitativă} \\ arg\ max_{1 \le \tau \le q} |\{ \kappa \mid \tilde{\varphi}_{z_k}(\bullet) = \tau, \kappa = 1 \div K\}|, \text{ dacă } Y \text{ este calitativă} \end{cases}$$

În primul caz, estimarea medie este media rezultatelor obţinute pentru modelele asociate fiecărui eşantion. În al doilea caz, a fost constituit un „comitet de modele" pentru a vota şi a alege răspunsul cel mai probabil. Când modelul returnează probabilităţi, asociate cu fiecare modalitate τ sau cu fiecare arbore de decizie, se calculează mediile acestor probabilităţi.

Practic, cele K eșantioane independente, $\boldsymbol{z}_\kappa$, ar necesita, în general, prea multe date și ele sunt înlocuite prin K eșantioane *bootsrap* , $\tilde{\boldsymbol{z}}_\kappa$, obținute, fiecare, prin n extrageri cu înlocuire conform legii empirice $\tilde{F}$. În fiecare iterație κ ($\kappa = 1 \div K$), se extrage eșantionul *bootstrap*, $\tilde{\boldsymbol{z}}_\kappa$

și se calculează $\tilde{\varphi}_{z\kappa}(\boldsymbol{x})$ pe acest eșantion. În final, după cum variabila de explicat Y este cantitativă sau calitativă, estimarea medie este sau media estimărilor sau rezultatul votului.

Păduri aleatoare. Pentru metoda segmentării o îmbunătățire a *bagging*-ului se poate obține prin adăugarea unei randomizări. Obiectivul este de mări independența arborilor de agregare prin intervenția hazardului în alegerea variabilelor implicate în modele. În fiecare iterație κ ($\kappa = 1 \div K$): se extrage un eșantion *bootstrap*, $\boldsymbol{z}^{\kappa}$ și se estimează un arbore pe $\boldsymbol{z}^{\kappa}$ prin randomizarea variabilelor (căutarea fiecărui nod optimal este precedată de selecția aleatoare a unei submulțimi de $q \leq p$ predictori). In final, $\tilde{\varphi}_{K}(\boldsymbol{x}) = (1/K)\sum_{\kappa=1}^{K}\tilde{\varphi}_{z\kappa}(\boldsymbol{x})$ sau $\tilde{\varphi}_{K}(\boldsymbol{x})$ = rezultatul votului. Față de *bagging*, în cazul „pădurilor aleatoare" de arbori decizionali (*Random Forest*), strategia de tăiere poate fi mai simplă limitându-se la arbori de mărimi, q, relativ reduse (chiar triviale: $q = 2$). Într-adevăr, doar cu *bagging* arborii limitați la o singură ramificație riscă să fie foarte asemănători (puternic corelați) implicând, aceleași, câteva variabile care apar ca fiind cele mai explicative. În fiecare etapă de construcție a unui arbore, selectarea aleatoare a unui număr redus de predictori potențiali crește semnificativ variabilitatea având în mod necesar alte variabile. Fiecare model de bază este în mod evident mai puțin eficient dar agregarea duce în cele din urmă la rezultate bune. Numărul de variabile extrase aleator nu este un parametru sensibil fapt pentru care Breiman (2001) sugerează alegerea implicită $q = p$. Evaluarea iterativă a erorii *out-of-bag* previne o eventuală supraajustare dacă aceasta tinde să se degradeze. Ca la toate modelele construite prin agregare (sau „cutie neagră"), pentru prospectorul de date nu există nici o interpretare directă. Informațiile relevante sunt obținute prin calcul și prin reprezentarea grafică a unor indici, proporționali cu importanța fiecărei variabile din modelul agregat adică cu participarea acesteia la regresie sau discriminare. Aceste informații sunt cu atât mai utile cu cât variabilele sunt mai numeroase. Pentru a evalua importanța unei variabile prospectorul de date utilizează criterii precum: frecvența cu care apare fiecare variabilă în arborii pădurii, *MDA* (*Mean Decrease Accuracy*) sau *MDG* (*Mean Decrease Gini*).

Strategii adaptive. Boosting-ul adoptă același principiu general ca și *bagging*-ul: construirea unei familii de modele care să fie agregate prin o medie ponderată a estimărilor sau a unui vot. El diferă net de *bagging* în ceeace privește modul de construire a familiei care, de această dată, este recurent: fiecare model este o versiune adaptivă a precedentului acordând, în momentul estimării următoare, o pondere mai mare observaţiilor prost ajustate sau prost previzionate. Intuitiv, acest algoritm își concentrează eforturile asupra observaţiilor celor mai dificil de ajustat astfel încât combinarea ansamblului de modele permite evitarea supraajustării.

Pentru exemplificare se consideră problema de discriminare în două clase și fie $\tilde{d}$ funcția de discriminare cu valori în {-1, 1}. Pentru estimarea primului model ponderile w_i ale fiecărei observații sunt inițializate la $1/n$, în continuare aceste ponderi evoluează la fiecare iterație adică pentru fiecare nouă estimare. Importanța, w_i, a unei observații rămâne neschimbată dacă observația este bine clasată, dacă nu este bine clasată w_i crește proporțional cu deficitul de ajustare al modelului. Agregarea finală a previziunilor, $\sum_{\kappa=1}^{K} c_{\kappa}\tilde{d}_{\kappa}(\boldsymbol{x})$, este o combinație ponderată a calităților de ajustare ale fiecărui model. Valoarea absolută a sa, numită marje, este proporțională cu încrederea care poate fi acordată semnului său care furnizează rezultatul previziunii.

Fie $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ un eșantion și $\boldsymbol{x}$ individul de previzionat. Se inițializează $\boldsymbol{w}_1$, vectorul de ponderi: $w_{1,i} = 1/n, \quad i = 1 \div n$. În fiecare iterație κ ($\kappa = 1 \div K$): se estimează $\tilde{d}_{\kappa}$ pe eșantionul

z_κ (z ponderat cu w_κ); se consideră vectorul $Q_\kappa = \{Q\kappa,i\}^n_{i=1}$, unde $Q\kappa,i$ este un indice de misclasare ($Q\kappa,i = 1$ dacă $\hat{d}_\kappa(x_i) \neq y_i$ și $Q\kappa,i = 0$ dacă $\hat{d}_\kappa(x_i) = y_i$); se estimează eroarea de previziune: $\tilde{\mathcal{E}}_P = (\sum^n_{i=1} w_i\, Q\kappa,i) / (\sum^n_{i=1} w_i)$; se calculează $c_\kappa = log(\ (1- \tilde{\mathcal{E}}_P) / \tilde{\mathcal{E}}_P\ )$; se calculează noile ponderi: $w_{\kappa+1,i} := w_{\kappa,i}\ exp[-c_\kappa Q\kappa,i]$, $i = 1 \div n$. În final, rezultatul votului este dat de formula: $\tilde{\varphi}_K(x) = sign[\sum^K_{\kappa=1} c_\kappa \hat{d}_\kappa(x)]$.

Principiile *bagging*-ului sau *boosting*-ului se pot aplica la orice metodă de modelare dar nu sunt interesante şi nu reduc sensibil eroarea de previziune decât în cazul modelelor instabile deci, mai degrabă, neliniare. Astfel, pentru prospectorul de date, utilizarea acestor algoritmi nu are nici un sens cu regresia multiliniară sau cu analiza discriminantă. Ei pot fi foarte utili în asociere cu arborii binari ca modele de bază.

Capitolul 3. ALIMENTAREA CU CUNOȘTINȚE A SISTEMELOR SUPORT PENTRU DECIZII

Rolul bibliotecilor în generarea/furnizarea de cunoștințe

Timp de secole, factorii de decizie au folosit conţinutul cărţilor, periodicelor, scrisorilor şi altor documente ca depozite textuale de cunoştinţe. Cunoştinţele încorporate într-un fragment de text pot fi descriptive, procedurale sau de raţionament. Indiferent de tipul acestora, factorii de decizie caută şi selectează piese de text pentru a dobândi mai multe cunoştinţe, pentru a verifica impresii sau pentru a stimula idei.

Bibliotecarii au început, prin anii ’70, să primească roluri decizionale active participând, în calitate de bibliotecari medicali clinici, la consultările pacienților unde, în funcție de diversele problemele identificate, formulau cu promptitudine căutări riguroase obținând rapid răspunsurile de actualitate cele mai utile echipelor medicale pentru luarea de decizii clinice consistent fundamentate. Sprijinul bibliotecilor și bibliotecarilor în luarea deciziilor a variat, în timp, de la unul pasiv (colecţii tradiţionale de cărţi şi reviste) către unele extrem de active (asistenți decizionali).

Generarea de cunoştinţe din texte a devenit posibilă și din ce în ce mai importantă prin funcţionalităţi precum *text-mining* sau *content analysis*. Generarea din hipertext a cunoştinţelor utile în procesele decizionale se realizează prin funcţionalităţi de tip *web-mining* (*web usage mining*, *web content mining* sau *web structure mining*).

Bibliotecile digitale au oferit perspective noi pentru sistemele suport pentru decizii ale companiilor. In societatea informațională, tot mai multe date digitale sunt colectate, procesate, gestionate și arhivate în biblioteci și centre de informare pentru a satisface, în fiecare moment, cerințele tot mai variate ale comunităților de utilizatori. Având în vedere imensitatea volumului de informații care se acumulează în bibliotecile digitale, unul dintre cei mai imperativi parametri de implementare a unui scenariu de extragere orientată către cerințe a informațiilor și de generare a cunoștințelor este *data mining*. Funcționalitățile *data mining* au devenit cruciale pentru gestionarea, organizarea informațiilor și diseminarea acestora către utilizatorii potriviți, la momentul potrivit.

Rezultatele explorării interconexiunilor dintre rețele sociale diferite au permis extinderea gamei de analize privind rețeaua constituită din comunitatea formată de autori și din comunitatea formată de bibliotecă împreună cu utilizatorii săi. *Bibliomining*, concept menit să susțină astfel de preocupări, a deschis perspectiva de a putea utiliza împreună, prin intermediul unui singur depozit de date, atât funcționalitățile oferite de bibliometrie cât și cele oferite de *data mining*.

Toate aceste evoluții precum și arhitectura generică a sistemelor suport pentru decizii, conturează și chiar susțin ideea, oportună și foarte tentantă, de a aborda construirea sistemelor suport pentru decizii ale bibliotecilor astfel încât acestea să poată oferi inclusiv funcționalitățile de alimentator de cunoștințe pentru alte sisteme decizionale ale unor, în special, mari companii.

Sistemul suport pentru decizii al unei biblioteci

Concepția și implementarea oricărui sistem informatic, deci și a unui sistem suport pentru decizii, sunt influențate de către o serie de factori, printre care pot fi menționați: obiectivele urmărite, evoluția mediului instituțional, normele și standardele utilizate, restricțiile impuse de către instituție, bugetul disponibil, persoanele implicate și termenele de finalizare.

Obiective. Provocările cu care se confruntă un sistem suport pentru decizii de bibliotecă sunt: elaborarea politicilor de achiziții și de diseminare orientate către cerere; optimizarea fluxurilor și alocării resurselor; îmbunătățirea conservării colecțiilor; diseminarea informațiilor către utilizatori; creșterea satisfacției utilizatorilor; comunicarea mai bună cu partenerii; diversificarea bunurilor culturale și creșterea veniturilor.

Principalele obiective ale sistemului sunt: extragerea, transformarea, încărcarea și integrarea datelor; simplificarea accesului la informații prin schimb transparent și diseminare accelerată a informațiilor; furnizarea de indicatori, de stare și de performanță, care să permită evaluarea în timp a conformității cu obiectivele bibliotecii; furnizarea de instrumente de analiză a tendințelor, de sesizare a situațiilor decizionale și de sugerare a unor acțiuni corespunzătoare în vederea fundamentării și luării deciziilor; asigurarea unor funcționalități de alimentator de cunoștințe pentru sistemele decizionale ale altor companii, interesate.

Arhitectură. Arhitectura sistemului se bazează pe combinarea tehnologiei de management a rezolvatoarelor flexibile cu tehnologia de management a bazelor de date. În Figura 2 este prezentată o variantă a acestei combinații, respectiv, integrarea depozitării datelor cu rezolvatoarele analitice (de prelucrare analitică *on-line*) şi cu rezolvatoarele *data mining* (de explorare a datelor şi descoperire a cunoştinţelor).

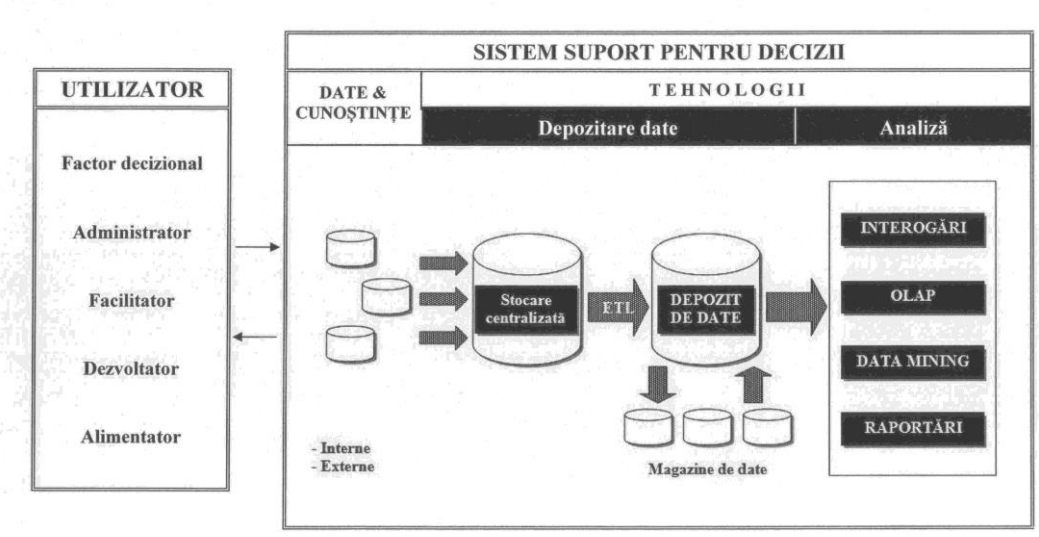


Figura XXX. Arhitectura sistemului suport pentru decizii al bibliotecii.

*Direcții de îmbunătățire a activităților***.** În funcție de cerințele principalelor categorii de utilizatori se evidențiază patru tipuri de servicii necesare: căutări simple care furnizează

rapoarte predefinite și indicatori de performanță; interogări avansate și/sau interogări personalizate, la cerere; căutări complexe, implicând navigare multidimensională și funcții puternice de analiză; simulări și/sau statistici sofisticate. Domeniile posibile de îmbunătățire a activității bibliotecii sunt: organizarea și conservarea colecţiilor; dezvoltarea colecţiilor; accesibilitatea colecțiilor; accesul la publicații; serviciile bibliografice, asistenţa și îndrumarea; utilizarea bibliotecii; digitalizarea colecțiilor; potenţialul de dezvoltare; managementul.

Avantaje. Pentru bibliotecă avantajele majore ale sistemului sunt: asigură instrumente performante de management și informații de calitate; rezolvă faze tehnice critice privind furnizarea, modelarea și stocarea datelor; satisface cerințele tehnice actuale și chiar viitoare; satisface cerințele utilizatorilor; adaptabilitatea; susține trecerea la cultura orientată către performanță și impune personalului dezvoltarea în consecință a abilităților. Pentru marile companii avantajele sistemului constau în asigurarea unor puternice funcționalități de alimentator de cunoștințe, pentru sistemele decizionale ale acestora, prin diseminarea informațiilor și cunoștințelor dorite de către utilizatorii interesați și la momente oportune. Procesul de realizare a unui astfel de sistem, nou și captivant, creează multe provocări dar promite mari îmbunătățiri în modul de desfășurare a activităților, în modul de înțelegere a ceea ce se face în prezent și a ceea ce se preconizează pentru viitor.

Variante de realizare. Având în vedere tipologiile deciziilor şi decidenţilor principalele modalități de realizare ale sistemelor suport pentru decizii ale bibliotecilor pot fi: sistem individual, folosit de o singură persoană pentru a-şi realiza propriile sarcini legate de elaborarea şi adoptarea deciziilor și destinat, în primul rând, decidenţilor individuali care lucrează independent; sistem colectiv, menit să asiste mai mulţi indivizi, cu poziţii de autoritate similare, care au de luat, în anumite momente, decizii colective; sistem instituțional, menit să faciliteze luarea acelor decizii care antrenează participanţi aflaţi pe niveluri ierarhice diferite; sistem orientat către comunicaţii, având drept componentă tehnlogică dominantă subsistemul de comunicaţii bazate pe calculator, menit să asiste codeciziile bazate pe comunicare şi colaborare între mai mulţi participanţi. În ceea ce privește bibliotecile, indiferent de forma de constituire şi administrare a patrimoniului (de drept public sau privat), acestea pot fi biblioteci clasice, biblioteci digitale și biblioteci mixte sau hibride.

Resurse necesare. Prezenta abordare vizează construirea unui sistem suport de decizii de nivel instituțional pentru o bibliotecă hibridă.

Prin resurse umane trebuie să se asigure următoarea structură de realizare: comitetul de management, echipa de proiect, grupurile de lucru cu utilizatorii, consultantul (firmă de consultanță pentru analiza cerințelor) și subcontractantul (firmă de specialitate pentru dezvoltare-implementare).

Prin resurse financiare trebuie să se asigure următoarea structură de produse și servicii: instrumentele de extracție, transformare și încărcare a datelor, instrumentele de raportare și diseminare, instrumentele de fundamentare a deciziilor, consultanța și dezvoltarea sistemului pe baze contractuale.

Prin resurse informaționale trebuie să se asigure următoarea structură de cerințe informaționale, dedusă din analiza obiectivelor instituționale: cerințele bibliografice, cerințele biblioteconomice și cerințele bibliometrice.

Analiza cerințelor informaționale

Cerințele bibliografice. Prin descriere bibliografică a unui document se înțelege o mulțime de informații privind patru aspecte (sau niveluri de analiză) diferite ale documentului descris și anume: *exemplarul*, conținând caracteristicile individuale ale unui singur exemplar al documentului; *manifestarea*, conținând caracteristicile publicației de care acesta aparține; *expresia*, conținând caracteristicile conținutului intelectual sau artistic și *lucrarea*, conținând caracteristicile creației abstracte la care se referă acest conținut. La fiecare din aceste patru niveluri de analiză, documentul descris este pus în relație cu o *persoană* sau cu o *colectivitate* care au intervenit într-un mod specific la nivelul respectiv.

Aceste șase noțiuni plus încă alte patru *loc*, *eveniment*, *obiect* și *concept* pot constitui subiecte ale unei lucrări și definesc entitățile esențiale, relevante pentru utilizatorii datelor bibliografice, grupate în: *produse* ale unei activități intelectuale sau artistice care sunt numite sau descrise în înregistrările bibliografice; *responsabilități* privind conținutul intelectual sau artistic, producția fizică, distribuția, gestionarea și aspectele juridice ale acestor produse; *subiecte* ale demersului intelectual sau artistic.

Id.	Denumire	Definire	Comentariu
	***Produse** ale unei activități intelectuale sau artistice*		
EPℓ	*lucrare*	o anumită creație/operă intelectuală sau artistică	entitate abstractă; identitatea de conținut a mai multor *expresii*
EPe	*expresie*	realizarea intelectuală sau artistică a unei *lucrări*	notație alfanumerică, muzicală coregrafică; formă sonoră, vizuală, obiectuală, cinetică
EPm	*manifestare*	materializarea unei *expresii* a unei *lucrări*	în functie de suport: manuscris, carte, periodic, afiș, film, casete, cd-uri
EPi	*exemplar*	un exemplar izolat al unei *manifestări*	un anumit exemplar al unei monografii, al unei casete audio, etc
	***Responsabili** pentru produsele unei activități intelectuale sau artistice*		
ERp	*persoană*	un individ	autor, compozitor, artist, editor, traducător, dirijor, interpret, etc
ERc	*colectivitate*	un organism sau un grup de indivizi/colectivități	identificat prin un nume specific și care acționează ca un tot
	***Subiecte** ale demersului intelectual sau artistic*		
ESκ	*concept*	o noțiune/idee abstractă	domeniu de cunoaștere, disciplină, teorie, metodă, tehnică, etc
ESω	*obiect*	o realitate materială	obiect natural/artificial, existent sau dispărut
ESε	*eveniment*	o acțiune sau un fapt	eveniment istoric, epocă, perioadă cronologică
ESλ	*loc*	date geografice/topografice	subiect al unei hărți, al unui atlas sau al unui ghid turistic

Tabelul 4. Entități bibliografice și semnificațiile lor.

Relațiile identificate între entitățile bibliografice reprezintă legăturile esențiale relevante pentru utilizatorii datelor bibliografice:

	Nume de relații dintre entități ↘	*lucrare*	*expresie*	*manifestare*	*exemplar*
1	*lucrare*	⇧↧ subiect ⇧↧ parte ⇧↧ succesor ⇧↧ supliment ⇧↧ complement ⇧↧ rezumat ⇧↧ adaptare ⇧↧ transformare ⇧↧ imitație			
2	*expresie*	↥⇩ subiect ↥⇩ realizare ↥⇩ succesor ↥⇩ supliment ↥⇩ complement ↥⇩ rezumat ↥⇩ adaptare ↥⇩ transformare ↥⇩ imitație	⇧↧ parte ⇧↧ succesor ⇧↧ supliment ⇧↧ complement ⇧↧ rezumat ⇧↧ adaptare ⇧↧ transformare ⇧↧ imitație ⇧↧ scurtare ⇧↧ revizuire ⇧↧ traducere ⇧↧ aranjament		
3	*manifestare*	↥⇩ subiect	⇧⇩ materializare	⇧↧ parte ⇧↧ reproducere ⇧↧ alternativă	
4	*exemplar*	↥⇩ subiect		↥⇩ reprezentare ↥⇩ reproducere	
5	*persoană*	⇧⇩ subiect	⇧⇩ realizare	⇧⇩ producere	⇧⇩ posesie
6	*colectivitate*	⇧⇩ subiect	⇧⇩ realizare	⇧⇩ producere	⇧⇩ posesie
7	*concept*	⇧⇩ subiect			
8	*obiect*	⇧⇩ subiect			
9	*eveniment*	⇧⇩ subiect			
10	*loc*	⇧⇩ subiect			
Notă: *Săgețile indică sensul și tipul fiecărei relații pe acel sens, respectiv 1:n* (⇨) sau *1:1* (↦)					

Tabelul 6. Complexitatea relațiilor dintre entitățile bibliografice.

Pentru entitățile bibliografice au fost identificate liste, considerate maximale, de atribute descriptive specifice necesare descrierilor bibliografice relevante ale acestora. În realitate, sursele de date pot oferi doar o parte din aceste informații, respectiv, surogate bibliografice. In cazul articolelor (documentelor) din revistele științifice (publicații) a rezultat:

	Ident	Nume atribut		Ident	Nume atribut		Ident	Nume atribut
Surogat bibliografic document								
1	*APℓ01*	*titlu-doc*	8	*APe08*	*volum-doc*	15	*APm17*	*restricții-doc*
2	*APℓ02*	*forma-doc*	9	*APe09*	*rezumat-doc*	16	*APm35*	*config-doc*
3	*APe02*	*tip-doc*	10	*APi06*	*stare-doc*	17	*APm36*	*fișier-doc*
4	*APℓ03*	*data-doc*	11	*APm02*	*resp-doc*	18	*APm37*	*acces-doc*
5	*APe04*	*limba-doc*	12	*APm03*	*ed-pub*	19	*APm38*	*adrURL-doc*
6	*APℓ06*	*domeniu-doc*	13	*APm13*	*format-doc*			
7	*APe05*	*subiect-doc*	14	*APm16*	*cost-doc*			
Surogat bibliografic publicație								
1	*APℓ01*	*titlu-pub*	10	*APℓ06*	*dom.-pub*	19	*APm15*	*furnizor-pub*
2	*APe10*	*context-pub*	11	*ASκ01*	*subiect-pub*	20	*APm16*	*cost-pub*
3	*APℓ02*	*forma-pub*	12	*APe08*	*volum-doc*	21	*APm17*	*restrict.-pub*
4	*APe02*	*tip-pub*	13	*APe09*	*rezumat-pub*	22	*APm22*	*stare-pub*
5	*APm14*	*id-int-pub*	14	*APm02*	*editor-pub*	23	*APm23*	*nrotare-pub*
6	*APℓ03*	*data-pub*	15	*APm05*	*editură-pub*	24	*APm35*	*config-pub*
7	*APe15*	*frecvența*	16	*APm03*	*ediție-pub*	25	*APm36*	*fișier-pub*
8	*APm04*	*țara*	17	*APm08*	*colecție-pub*	26	*APm37*	*acces-pub*
9	*APe04*	*limba-pub*	18	*APm13*	*format-pub*	27	*APm38*	*adrURL-pub*

Tabelele 17-18. Maparea pe sursele de date a descrierilor bibliografice

Cerințe biblioteconomice. Procesele biblioteconomice sunt văzute ca succesiuni de activități formate la rândul lor din secvențe de operații elementare consumatoare de resurse. O operație elementară, $\theta \in \Theta$, este descrisă într-un nomenclator Θ, specific unei anumite activități, prin elemente descriptive precum: durata, $\tau(\theta)$; cantitatea, $q(\theta)$; costul, $c(\theta)$; termenul de realizare, $t(\theta)$; etc.

Indicele de selecție al unei operații, θ, este o valoare scalară $s(\theta) \in \{0, 1\}$ care descrie faptul că, în conformitate cu o anumită politică de planificare/selecție, pentru operația analizată θ, descrisă în nomenclatorul de operații Θ, se consideră necesară efectuarea ei:

$$s(\theta,\Theta) = \begin{cases} 1 & \text{dacă efectuarea operației } \theta \in \Theta \text{ este considerată necesară;} \\ 0 & \text{în caz contrar} \end{cases}$$

Indicele de realizare al unei operații este o valoare scalară, $r(\theta) \in \{0, 1\}$, care descrie faptul că operația analizată θ, descrisă în nomenclatorul de operații Θ, a fost realizată:

$$s(\theta,\Theta) = \begin{cases} 1 & \text{dacă operația } \theta \in \Theta \text{ a fost realizată;} \\ 0 & \text{în caz contrar} \end{cases}$$

In procesele biblioteconomice curente se realizează și operații care nu fac obiectul unei politici de selecție dar care pot face obiectul unor solicitări aleatoare ale utilizatorilor.

Funcția de selecție este o funcție $S(\Theta, \bullet) : \Theta \times N \rightarrow N$, unde $S(\Theta, t)$ este o valoare scalară care reprezintă numărul tuturor operațiilor θ selectate din nomenclatorul Θ al activității analizate pentru a fi efectuate înainte de momentul t. O variantă, simplă dar evaluabilă, de

definire a funcției de selecție pentru activitatea analizată Θ și pentru intervalul de timp analizat, $T = [0, t]$ este: $S(\Theta, T) = \sum_{\theta \in O(\Theta, T)} s(\theta, \Theta)$, unde $O(\Theta, T) = \{ \theta \mid \theta \in \Theta, t(\theta) \in T \}$.

Funcția de realizare este o funcție $R(\Theta, \bullet) : \Theta \times N \rightarrow N$, unde $R(\Theta, t)$ este o valoare scalară care reprezintă numărul tuturor operațiilor θ din nomenclatorul Θ al activității analizate realizate înainte de momentul t. O variantă, simplă și evaluabilă, de definire a funcției de realizare pentru activitatea analizată Θ și intervalul de timp analizat $T = [0, t]$ este: $R(\Theta, T) = \sum_{\theta \in O(\Theta, T)} r(\theta, \Theta)$, unde $O(\Theta, T) = \{ \theta \mid \theta \in \Theta, t(\theta) \in T \}$.

Indicii și respectiv funcțiile de realizare și/sau de selecție permit prin modalități de agregare specifice obținerea valorilor tuturor indicatorilor operaționali și de performanță ai bibliotecii pe intervalul de timp analizat furnizându-se astfel descrierea stării curente și/sau dorite a sistemului instituției precum și evaluarea în timp a conformității cu obiectivele bibliotecii. De exemplu, în cazul în care Θ reprezintă lista tuturor titlurilor deținute de bibliotecă pentru activitatea de împrumut individual, o operație elementară $\theta \in \Theta$ vizează împrumutul unui singur titlu. Notând cu $O(\Theta, T) = \{ \theta \mid \theta \in \Theta, t(\theta) \in T \}$ mulțimea titlurilor solicitate de către utilizatori, pentru împrumut individual, în intervalul de timp analizat $T = [0, t]$ se obțin formule de definire evaluabile pentru doi dintre indicatorii operaționali ai unei biblioteci precum și pentru un indicator de performanță:

$\boldsymbol{S} \equiv S(\Theta, T) = \sum_{\theta \in O(\Theta, T)} s(\theta, \Theta)$ = numărul total de Titluri solicitate de către utilizatori prin împrumut individual;

$\boldsymbol{s} \equiv R(\Theta, T) = \sum_{\theta \in O(\Theta, T)} r(\theta, \Theta)$ = numărul total de Titluri servite către utilizatori prin împrumut individual;

$\boldsymbol{P} \equiv P(\Theta, T) = (\boldsymbol{s} / \boldsymbol{S}) \times 100$ = ponderea titlurilor deținute de bibliotecă în numărul total de titluri solicitate de către utilizatori.

Cerințe bibliometrice. Indicatorii destinați pentru a măsura productivitatea cercetătorilor sau a grupurilor de cercetare sunt considerați indicatori cantitativi.

Indicele de publicare al unui document, $\pi(d, x')$, este o valoare scalară care descrie faptul că una sau mai multe entități $x' \in X'(d)$, $X'(d) \subset X$ au contribuit în mod specific la publicarea documentului $d \in D$:

$$\pi(d, x') = \begin{cases} 1 & \text{dacă } x' \in X'(d) \\ 0 & \text{în caz contrar} \end{cases}$$

Funcția de publicare este o funcție $\Pi(x, \bullet)$: $X \times N \rightarrow N$, unde $\Pi(x, t)$ este o valoare scalară care reprezintă numărul tuturor publicărilor produse de entitatea analizată, $x \in X$, înainte de momentul (anul) t. O variantă, simplă și evaluabilă, de definire a funcției de publicare pentru entitatea analizată x și intervalul de timp analizat $T = [0, t\text{-}1]$ este: $\Pi(x, t) = \sum_{d \in D(x, T)} \pi(d, x)$, unde $D(x, T) = \{ d \mid d \in D(x), t(d) \in T \}$.

Indicatorii care ajută la identificarea nivelului de calitate al lucrărilor unui cercetător sau ale unui grup de cercetare și care pot fi utilizați pentru a evalua impactul cercetărilor în comunitatea științifică sunt considerați indicatori de performanță.

Indicele de impact al unui document este o valoare scalară care descrie faptul că un anumit document $d \in D$ a fost citat într-un un alt document $d' \in D$, $d' \neq d$:

$$\rho(d, d') = \begin{cases} 1 & \text{dacă pentru } d \text{ există o referință în } d' \\ 0 & \text{în caz contrar} \end{cases}$$

Indicele de notorietate al unei entități analizate, x, este un scor $\varepsilon(x)$ atașat lui x de către experți, membri ai unor centre recunoscute ca autorități științifice.

Indicele de încredere al unui document, d, este un indice $\varepsilon(d)$ care depinde de toți sau de o parte a indicilor de notorietate atașați entităților care sunt considerate în relație cu acel document: $\varepsilon(d) = \phi(\varepsilon(A(d)), \varepsilon(E(d)), \varepsilon(P(d)), \varepsilon(G(A(d))))$ respectiv autorul, editura, publicația sau grupul la care este afiliat autorul. O variantă evaluabilă, de definire a indicelui de încredere al lui d, este: $\varepsilon(d) = (w_A \varepsilon(A(d)) + w_E \varepsilon(E(d)) + w_P \varepsilon(P(d)) + w_G \varepsilon(G(A(d)))) / \mathcal{E}$, unde: $w_A + w_E + w_P + w_G = 1$ cu $w_A, w_E, w_P, w_G \geq 0$ și $\mathcal{E} = \varepsilon(A(d)) + \varepsilon(E(d)) + \varepsilon(P(d)) + \varepsilon(G(A(d)))$. $\varepsilon(d)$ este un indice à priori, care descrie un document d în momentul publicării, înainte de a se obține informații despre referințele la d.

Indicele de relevanță al unei citări este o valoare scalară, $\sigma(d, d') \geq 0$, care descrie cât de relevantă poate fi considerată citarea documentului $d \in D$ de către documentul $d' \in D$:

$$\sigma(d, d') = \begin{cases} > 0 & \text{dacă } d \text{ este citat în } d' \\ 0 & \text{în caz contrar} \end{cases}$$

O formulă evaluabilă pentru relevanța citării lui d de către d' este: $\sigma(d, d') = M / (m + M)$, unde: $d \in D(a)$, $d' \in D(a')$, $M = max\{\rho(a, a'), \rho(a', a)\}$ și $m = min\{ \rho(a, a'), \rho(a', a)\}$; avem $\sigma(d, d') \in [0.5, 1]$; dacă $a \neq a'$ atunci m reprezintă numărul de citări reciproce iar dacă $a = a'$ atunci m reprezintă numărul de autocitări.

Funcția de impact a unui document, d, este funcția $I(d, \bullet) : D \times N \rightarrow \mathcal{R}^+$, unde $I(d, t)$ este o valoare scalară care descrie impactul tuturor referințelor la documentul $d \in D$ înainte de momentul (anul) t. $I(d, t)$, valoarea funcției de impact a lui d la momentul t, depinde de: numărul $\rho(d)$ de citări ale documentului d în intervalul de timp $T = [t(d), t-1]$ unde $t(d)$ este anul publicării documentului d și de indicii $\varepsilon(d')$ și $\sigma(d, d')$ care descriu credibilitatea documentelor d' care citează pe d și respectiv relevanța acestor citări. O variantă, calculabilă, de definire a funcției de impact a unui document analizat, d, este: $I(d, t) = \sum_{d' \in D(T)} \rho(d, d')$, unde: $T = [t(d), t-1]$ este intervalul de timp analizat; sumarea se face pentru toate documentele d' care conțin o referință la d și au fost publicate în intervalul de timp T, $t(d') \in T$.

Funcția de impact a unei entități analizate, x, pentru o fereastră de citare de n ani, este este o funcție $I_n(x, \bullet) : X \times N \rightarrow \mathcal{R}^+$, unde $I_n(x, t)$ este o valoare scalară care descrie impactul din momentul t al tuturor documentelor publicate de entitatea analizată, x, într-un interval de timp analizat, $T = [t-n, t-1]$: $I_n(x, t) = \sum_{d \in D(x, T)} I(d, t)$, unde $I(d, t)$ este valoarea funcției de impact a documentului d la momentul t; sumarea se face pentru toate documentele d publicate de entitatea x în intervalul de timp analizat, $t(d) \in T$.

Factorul de impact al unei entități analizate, x, pentru o fereastră de citare de n ani, este: $IF_n(x) = I_n(x, t) / \Pi(x, T)$, unde $T = [t-n, t-1]$ este intervalul de timp analizat (fereastra de citare); $I_n(x, t)$ este valoarea din anul t a funcției de impact a entității x pentru perioada T iar $\Pi(x, T)$ reprezintă numărul total de documente publicate de entitatea x în aceeași perioadă.

Indicele de notorietate al unei mulțimi de documente, X, este un indice $\mathcal{E}(X)$ care depinde de indicii de notorietate ai editurilor și/sau publicațiilor pentru fiecare $d \in X$. În mod obișnuit avem $X = D(x)$ unde entitatea analizată x poate fi un autor a, un grup de cercetare g, o publicație p sau o editură e: $\mathcal{E}(X) \equiv \mathcal{E}(D(x)) = \phi (\{(\varepsilon(E(d)), \varepsilon(P(d))) \mid d \in D(x)\})$. O variantă calculabilă a definiției este: $\mathcal{E}(D(x)) = \Sigma_{d \in D(x)} (w_E(\varepsilon(E(d)) + w_P \varepsilon(P(d)))$, unde $w_E + w_P = 1$ cu $w_E, w_P \geq 0$.

Indicele de notorietate al unui autor, a, este un indice $\mathcal{E}_3(a)$ care depinde de a și de afilierea acestuia, $G(a)$. O variantă evaluabilă a definiției este: $\mathcal{E}(a) = w_A \varepsilon(a) + w_G \varepsilon(G(a))$, unde $w_A + w_G = 1$ cu $w_A, w_G \geq 0$.

Indicele de notorietate-impact al unei mulțimi de documente, X, este un indicator $\mathcal{EI}(X)$ care depinde de indicele de notorietate $\mathcal{E}(X)$ și de valoarea funcției de impact $I(X, t)$, în anul de

referință *t*. Cea mai simplă formă de definiție calculabilă este: $\mathcal{EI}(X) := w_1\, \mathcal{E}(X) + w_2\, I(X, t)$, unde $w_1 + w_2 = 1$ cu $w_1, w_2 \geq 0$.

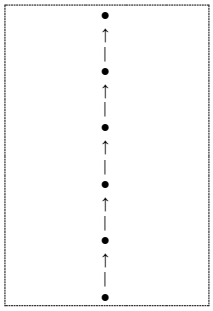
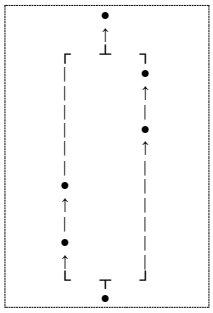
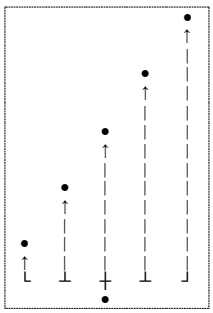
Noțiunile definite au permis obținerea de definiții evaluabile, pentru oricare dintre indicatorii bibliometrici uzuali, în strictă concordanță semnificațiile curente ale acestor indicatori.

Depozitarea datelor

Identificare fapte. Pentru mediul decizional al unei biblioteci subiectele majore de interes identificate sunt: serviciile de bibliotecă, aparițiile editoriale și calitatea publicațiilor.

Definire dimensiuni. Perspectivele de analiză, necesare mediului decizional, pentru fiecare din faptele identificate sunt: timp, operație și utilizator pentru serviciile de bibliotecă; timp, publicare, autor, editor, publicație și subiect pentru aparițiile editoriale precum și timp, publicare, autor, referință, publicație și subiect pentru calitatea publicațiilor. Schema dimensională a depozitului de date este prezentată în Figura 15.

Schema dimensională a depozitului de date

Dimens	*Niveluri*	*Căi de agregare*	*Descrieri*
$D \in \boldsymbol{D}$			
1 TIMP	perioadă		-♦ anii
	an		-♦ anul
	semestru		-♦ semestrul
	trimestru		-♦ trimestrul
	lună		-♦ luna
	zi		-♦ ziua din lună
2 OPERAȚIE	sistem		-♦ instituția
	proces		-♦ procesul
	activitate		-♦ activitatea
	compartiment		-♦ denumirea
	post		-♦ angajatul -♦ funcția
	operație		-♦ denumirea -♦ codul
3 UTILIZATOR	continuitate		-♦ re/nou înscris
	naționalitate		-♦română/altă
	gen		-♦ masc./fem.
	vârstă		-♦ categoria
	ocupație		-♦ statutul
	utilizator		-♦ numele -♦ permisul

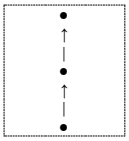
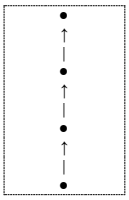
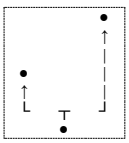
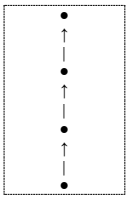
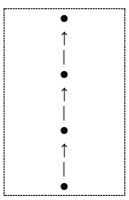
Dimens	*Niveluri*	*Căi de agregare*	*Descrieri*
4 SUBIECT	domeniu		-♦ denumirea
	subdomeniu		-♦ denumirea -♦ descriptorii
	subiect		-♦ denumirea -♦ descriptorii
5 PUBLICAȚIE	țară		-♦ denumirea
	oraș		-♦ denumirea
	editură		-♦ denumirea -♦ adresa
	publicație		-♦ titlul -♦ frecvența -♦ limba
6 DOCUMENT	format		-♦ denumirea
	tip		-♦ denumirea
	document		-♦ titlul -♦ limba
7 AUTOR	țară		-♦ denumirea
	oraș		-♦ denumirea
	afiliere		-♦ instituția -♦ adresa
	autor		-♦ numele -♦ profesia -♦ adresa
8 EDITOR	țară		-♦ denumirea
	oraș		-♦ denumirea
	afiliere		-♦ instituția -♦ adresa
	autor		-♦ numele -♦ profesia -♦ adresa

Figura 15. Schema dimensională a depozitului de date

Definire măsuri. Aspectele specifice și măsurabile ale faptelor, relevante pentru analiză, la nivelul minim de granularitate, sunt: indicii de selecție (s) și de realizare (r), duratele (τ) și costurile (c) unitare ale operațiilor pentru serviciile de bibliotecă; indicii de publicare (π) și de cotare (ψ) pentru aparițiile editoriale precum și indicii de notorietate (ε), de impact (ρ) și de relevanță a citărilor (σ) pentru calitatea publicațiilor.

Setul de interogări preliminare. Sistemele de indicatori (operaționali, de performanță și bibliometrici), definiți anterior, reprezintă de fapt cerințe ale utilizatorilor și constituie setul de interogări preliminare la care depozitul de date poate răspunde.

Modelul multidimensional al depozitului de date. Etapa de modelare multidimensională a datelor, fundamentată pe analiza cerințelor informaționale deduse din obiectivele instituționale și pe reconcilierea cu sursele de date, a permis identificarea faptelor, definirea dimensiunilor, nivelurilor dimensionale, măsurilor, căilor de agregare și arborilor de atribute , respectiv, cuburile de date.

Schemele cuburilor de date sunt reprezentate prin diagrame în care faptele sunt reprezentate prin dreptunghiuri, dimensiunile sunt reprezentate prin dreptunghiuri rotunjite iar măsurile sunt reprezentate prin cercuri.

Cubul de date „Servicii bibliotecare”

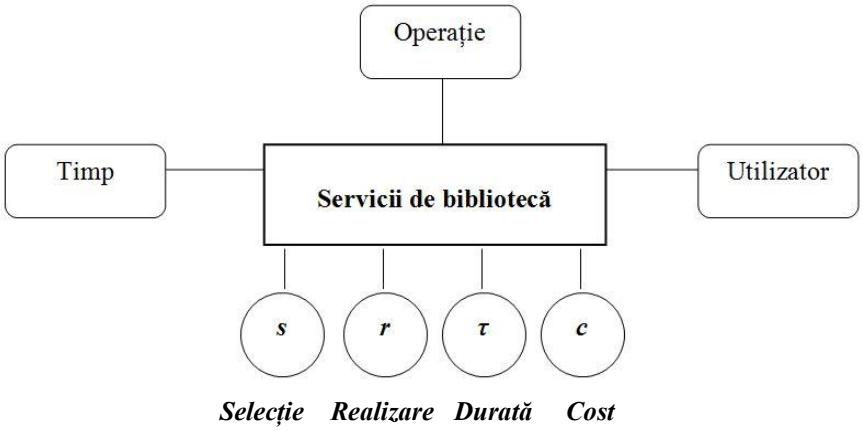


Figura 16. Cub de date privind serviciile bibliotecare

Cubul de date „Apariții editoriale”

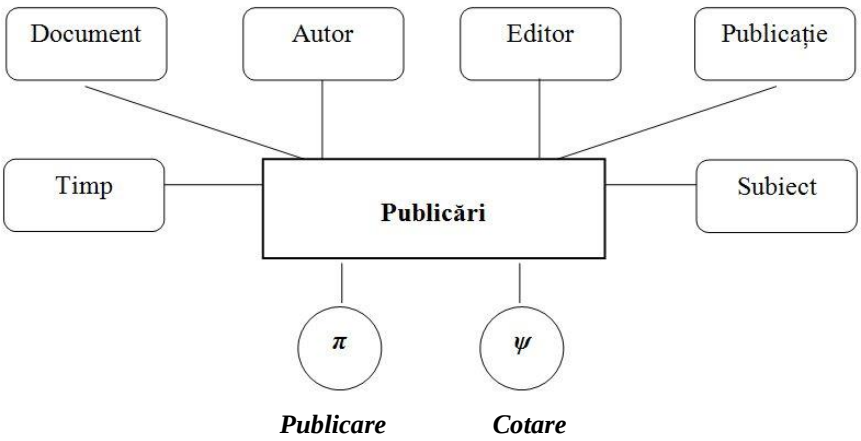


Figura 17. Cub de date privind aparițiile editoriale

Cubul de date „Calitatea publicărilor”

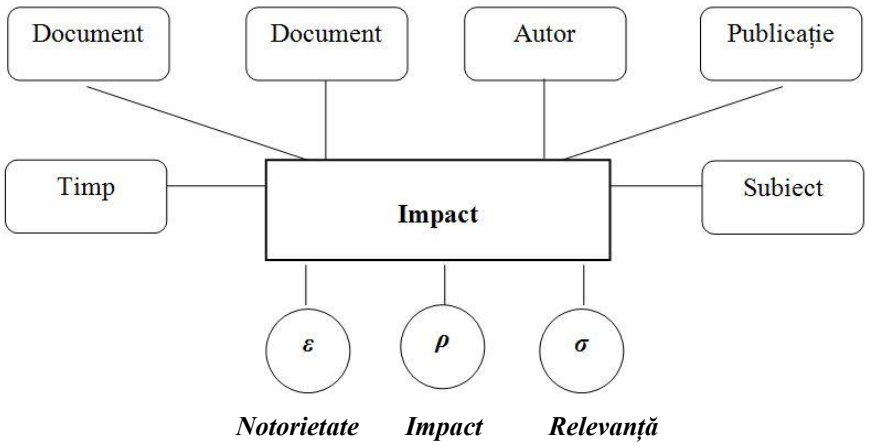


Figura 18. Cub de date privind calitatea publicărilor

Schema conceptuală a depozitului de date.

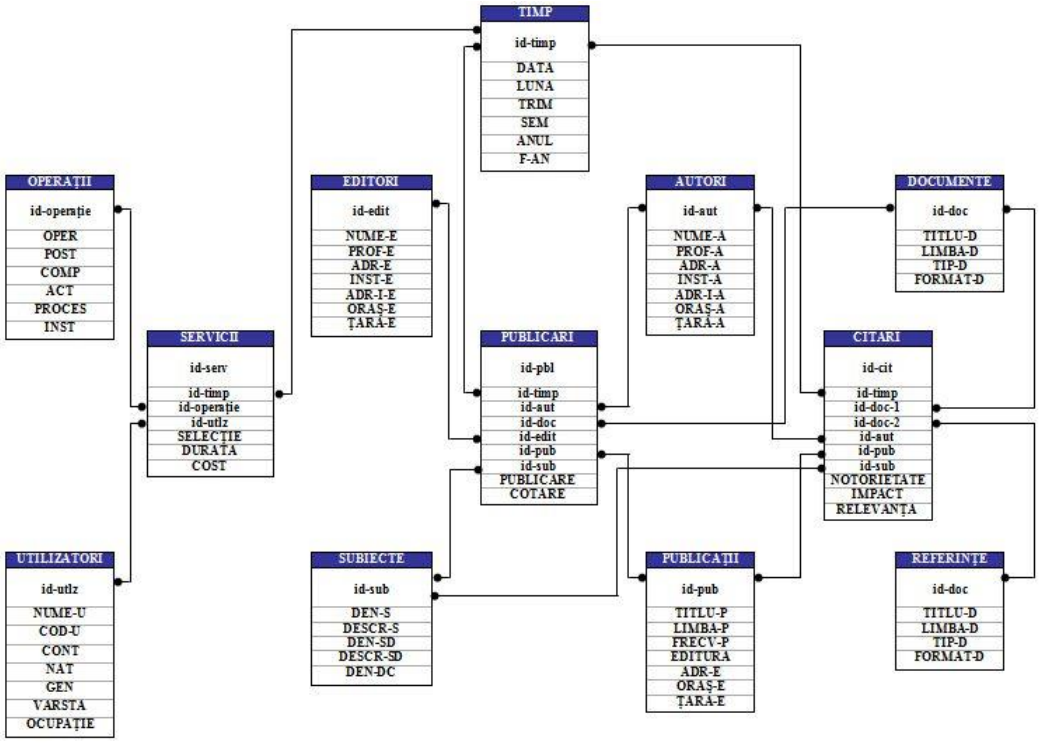


Figura 19. Schema conceptuală („constelație”) a depozitului de date.

Descoperirea/generarea de cunoștințe din (depozitul de) date

Printre problemele de referință din sistemele decizionale ale bibliotecilor, rezolvabile prin *data mining*, se pot evidenția: identificarea de nuclee de autoritate în diferite mulțimi de entități, analiza grupurilor de entități, elaborarea de recomandări.

Au fost definite proceduri privind ierarhizarea preferințelor de lectură ale utilizatorilor, ierarhizarea subiectelor în raport cu interesul utilizatorilor, ierarhizarea autorilor care tratatează un anumit subiect, gruparea bazată pe conținut a documentelor și recomandarea

către utilizatori a documentelor intrate recent în colecțiile bibliotecii. Astfel de proceduri pot fi adaptate foarte ușor și pentru alte entități publicații, edituri sau grupuri științifice.

Modelul conceptual al depozitului de date descrie datele multidimensionale independent de implementarea (logică) particulară. Cuburile de date sunt reprezentate (grafic) prin tabele, această reprezentare sugerează cum pot fi implementate cuburile de date cu ajutorul modelului relaţional. Setul de date de test, respectiv, setul de instanțieri ale schemei multidimensionale, a fost creat și administrat utilizând sistemul Access. Pentru verificarea funcționalității procedurilor s-a realizat un sistem de module de test. Rezultatele experimentale obținute au fost atașate fiecărei proceduri și sunt menite să ilustreze, în special, modurile de desfășurare ale proceselor computaționale.

Ierarhizarea preferințelor utilizatorilor. Ierarhizarea preferințelor de lectură ale utilizatorilor bibliotecii presupune identificarea acelor documente care au fost consultate împreună în mod frecvent. Identificarea se dorește să fie făcută, în mod automat, pe baza operațiilor de împrumut pentru lectură $O(\Theta^{\ell}, T)$ realizate în perioada de timp analizată, T. Pentru fiecare operație, θ, există înregistrate în depozitul de date: documentul consultat, $d \in \mathcal{D}$, utilizatorul care a realizat consultarea, $u \in U$ și momentul realizării consultării, $t \in T$,($[\theta] = [d, u, t]$). Se consideră că două documente diferite, $d' \neq d''$, au fost consultate împreună dacă ele au fost consultate de același utilizator în aceeași unitate de timp adică dacă operațiile au fost simultane: $\theta' \bowtie \theta'' \Leftrightarrow (d' \neq d'') \wedge (u' = u'') \wedge (t' = t'')$. Setul de documente definit de mai multe operații simultane $\theta' \bowtie \theta'' \bowtie \ldots$ formează o tranzacție admisibilă, $\boldsymbol{\theta} = \{d', d'', \ldots\} \in \boldsymbol{\Theta}$.

Fie $\boldsymbol{D} = \{d_j\}^{m}_{j=1}$ mulțimea documentelor consultate împreună, fie $\boldsymbol{\Theta} = \{\boldsymbol{\theta_i}\}^{n}_{i=1}$ mulțimea tranzacțiilor admisibile cu documentele din $\boldsymbol{D}$, $\boldsymbol{\theta_i} \subseteq \boldsymbol{D}$ și fie $\boldsymbol{x} \subseteq \boldsymbol{D}$. Mulțimea de tranzacții admisibile $\boldsymbol{\theta_i}$ din $\boldsymbol{\Theta}$ care îl conțin pe $\boldsymbol{x}$, $\mathcal{K}_{\boldsymbol{\Theta}}(\boldsymbol{x}) = \{ \boldsymbol{\theta_\ell} \mid \boldsymbol{\theta_\ell} \supseteq \boldsymbol{x}, \ell \in [1, n] \}$, este numită acoperire a lui $\boldsymbol{x}$ în $\boldsymbol{\Theta}$. Contorul de suport al lui $\boldsymbol{x}$ este numărul de tranzacții care îl conțin pe $\boldsymbol{x}$: $\boldsymbol{x_\sigma} = |\mathcal{K}_{\boldsymbol{\Theta}}(\boldsymbol{x})|$. Se consideră că $\boldsymbol{x}$ este (consultat) frecvent dacă $\boldsymbol{x_\sigma} \geq \sigma_{min}$ unde σ_{min} reprezintă un prag de suport ales de către utilizator.

Prin scanarea mulțimii $O(\Theta^{\ell}, T)$ de operații de consultare realizate în perioada de timp analizată, procedura generează, mai întâi, mulțimea de documente consultate simultan. Se obține drept rezultat lista $\boldsymbol{D}$ a *setdoc*-urilor $\boldsymbol{x}$ de dimensiune $dim(\boldsymbol{x}) = 1$, respectiv secțiunea de documente analizate. Procedura continuă prin constituirea de tranzacții admisibile candidate pentru fiecare utilizator și fiecare unitate de timp c_{ut}. Ulterior se determină contorul, intermediar, de suport la nivelul fiecărui utilizator u_σ și contorul de suport c_σ pentru fiecare tranzacție. Ierarhizarea dorită se obține prin furnizarea tuturor tranzacțiilor candidate în ordinea descrescătoare a valorilor contorului de suport c_σ. În final, sunt reținute tranzacțiile pentru care este respectat pragul de suport, $c_\sigma \geq \sigma_{min}$.

Ierarhia consultărilor simultane (frecvente și nefrecvente)

Nr. crt	$dim(\boldsymbol{\theta})$	$\boldsymbol{\theta} = \{d_\kappa \mid \kappa \in K\}$				$id(\kappa,\boldsymbol{\theta})$		Nr. crt	$dim(\boldsymbol{\theta})$	$\boldsymbol{\theta} = \{d_\kappa \mid \kappa \in K\}$				$id(\kappa,\boldsymbol{\theta})$	
	κ	d_1	d_2	d_3	d_4	c	c_σ		κ	d_1	d_2	d_3	d_4	c	c_σ
1	2	1	5			15	159	**21**	2	11	14			28	112
2	3	7	9	11		42	152	**22**	3	1	2	6		34	112
3	2	7	14			25	149	**23**	3	1	2	4		32	107
4	2	1	6			16	146	**24**	2	7	9			23	105
5	3	2	4	5		38	146	**25**	2	1	4			14	103
6	4	7	9	11	14	52	142	**26**	3	7	11	14		44	102
7	4	1	4	5	6	50	140	**27**	2	5	6			22	99
8	3	1	5	6		37	137	**28**	2	18	20			30	99
9	4	1	2	4	6	48	134	**29**	3	1	2	5		33	99
10	3	18	19	20		46	133	**30**	3	7	9	14		43	97
11	2	1	2			13	132	**31**	2	2	5			18	92

12	*2*	*9*	*14*			*27*	*130*	***32***	*2*	*9*	*11*			*26*	*92*
13	*2*	*18*	*19*			*29*	*127*	***33***	*2*	*2*	*6*			*19*	*91*
14	*2*	*19*	*20*			*31*	*127*	***34***	*2*	*7*	*11*			*24*	*90*
15	*3*	*9*	*11*	*14*		*45*	*126*	*35*	*3*	*2*	*4*	*6*		***39***	***80***
16	*2*	*4*	*6*			*21*	*125*	*36*	*3*	*1*	*4*	*5*		***35***	***74***
17	*4*	*2*	*4*	*5*	*6*	*51*	*121*	*37*	*3*	*4*	*5*	*6*		***41***	***64***
18	*2*	*2*	*4*			*17*	*118*	*38*	*2*	*4*	*5*			***20***	***62***
19	*4*	*1*	*2*	*4*	*5*	*47*	*117*	*39*	*3*	*2*	*5*	*6*		***40***	***49***
20	*4*	*1*	*2*	*5*	*6*	*49*	*113*	*40*	*3*	*1*	*4*	*6*		***36***	***31***

Tabelul 23. Ierarhia tranzacțiilor frecvente

Ierarhizarea subiectelor de interes. Fie $c \in C$ (sub)domeniul de cercetare analizat, fie $S(c)$ mulțimea de subiecte din acest domeniu de interes și fie $O(\Theta, T)$ mulțimea tuturor operațiilor de consultare de documente realizate în perioada de timp analizată, T. Se dorește o ierarhizare în interiorul mulțimii $S(c)$ pe perioada T. Ierarhizarea presupune identificarea automată a subiectelor de interes abordate în cadrul fiecărui document consultat, transformarea operațiilor de consultare-document în operații de consultare-subiect și contorizarea acestora pe fiecare subiect în parte.

In general, într-un document, d, sunt abordate mai multe subiecte, $S(d)$, astfel încât, este firesc să se presupună că prin consultarea documentului, au fost consultate (implicit) toate subiectele abordate în acel document. Accesarea, de către un utilizator, u, a unui document, d, poate fi realizată prin mai multe tipuri specifice de operații $\theta^{\kappa} \in \Theta^{\kappa}$, $\kappa \in \mathcal{K}$. Pentru fiecare astfel de operație există înregistrate în depozitul de date: tipul de accesare, $\kappa \in \mathcal{K}$, documentul consultat, $d \in \mathcal{D}$, utilizatorul care a realizat consultarea, $u \in U$ și momentul realizării consultării, $t \in T$, ($[\theta^{\kappa}] = [\kappa, d, u, t]$). De asemenea, pentru fiecare document, $d \in \mathcal{D}$, există înregistrate în depozitul de date toate subiectele abordate în documentul respectiv, $S(d)$. În acest context, interesează doar $S(d, c) = S(d) \cap S(c)$ adică numai acele subiecte abordate în d care aparțin domeniului analizat.

Procedura generează, mai întâi, secțiunea de documente analizate, respectiv, mulțimea tuturor documentelor consultate în perioada analizată care abordează subiectele de interes: $D(S(c)) = \cup_{s \in S(c)} D(s)$. În continuare, prin scanarea mulțimilor de operații de consultare document, $O(\Theta^{\kappa}, T)$, pentru fiecare operație, $[\theta] = [\kappa, d, u, t]$, se generează câte un set de operații de consultare subiect, $\{[\kappa, s, d, u, t]\}_{s \in S(d, c)}$, corespunzător subiectelor de interes abordate în documentul consultat, d.

Se consideră că, pentru orice document, $d \in \mathcal{D}$, valoarea funcției de realizare a operației de consultare a unui subiect de interes, $s \in S(d, c)$, este dată de valoarea funcției de realizare a operației de consultare a documentului d : $R_s(\Theta^{\kappa}, T) = R_d(\Theta^{\kappa}, T)$. Pentru fiecare subiect, $s \in S(c)$, valoarea funcției de realizare a operației de consultare a subiectului se obține prin cumularea valorilor funcțiilor de realizare ale operațiilor de consultare document, pentru toate documentele și toate tipurile de operații de consultare. În final, procedura furnizează, în ordine descrescătoare, valorile funcțiilor de realizare pentru operațiile de consultare subiect obținute în perioada de timp analizată, T, pe fiecare subiect de interes.

Ierarhia subiectelor de interes $s_i \in S(c)$								
$d \in \mathcal{D}$	$R_d(\Theta^{\kappa},T)$	$\kappa \in \mathcal{K}$	s_1	s_4	s_5	s_6	s_3	s_2
84	*688*	*b*	*688*	*688*	*688*		*688*	*688*
89	*834*	*b*	*834*		*834*	*834*	*834*	*834*
91	*1.116*	*b*		*1.116*	*1.116*	*1.116*		*1.116*
93	*1.679*	*b*	*1.679*	*1.679*		*1.679*		*1.679*
96	*1.016*	*b*	*1.016*	*1.016*			*1.016*	*1.016*
97	*1.265*	*b*	*1.265*		*1.265*	*1.265*		*1.265*
99	*571*	*b*		*571*	*571*		*571*	*571*
42	*1.698*	*e*	*1.698*				*1.698*	

43	*701*	*e*	*701*	*701*	*701*	*701*		*701*
44	*616*	*e*	*616*	*616*			*616*	
45	*525*	*e*			*525*	*525*	*525*	*525*
48	*368*	*e*		*368*	*368*	*368*	*368*	
50	*1.176*	*e*	*1.176*	*1.176*		*1.176*	*1.176*	
53	*1.312*	*e*	*1.312*		*1.312*	*1.312*		
54	*1.605*	*e*	*1.605*	*1.605*	*1.605*			*1.605*
55	*356*	*e*	*356*	*356*		*356*	*356*	
56	*1.667*	*e*		*1.667*	*1.667*		*1.667*	
58	*648*	*e*			*648*	*648*	*648*	*648*
59	*48*	*e*			*48*	*48*		*48*
64	*1.144*	*e*		*1.144*				*1.144*
66	*71*	*e*		*71*	*71*	*71*		*71*
69	*404*	*e*	*404*	*404*	*404*	*404*	*404*	
71	*250*	*e*	*250*	*250*	*250*		*250*	*250*
72	*494*	*e*	*494*	*494*	*494*		*494*	
73	*486*	*e*	*486*	*486*	*486*	*486*	*486*	*486*
75	*1.241*	*e*		*1.241*		*1.241*		*1.241*
77	*1.013*	*e*	*1.013*	*1.013*	*1.013*	*1.013*	*1.013*	*1.013*
80	*477*	*e*				*477*		*477*
1	*1.275*	*ℓ*						
2	*313*	*ℓ*	*313*					*313*
4	*1.190*	*ℓ*		*1.190*	*1.190*		*1.190*	
5	*1.162*	*ℓ*	*1.162*		*1.162*	*1.162*		*1.162*
6	*1.658*	*ℓ*	*1.658*		*1.658*	*1.658*	*1.658*	*1.658*
7	*704*	*ℓ*	*704*	*704*	*704*	*704*	*704*	
9	*977*	*ℓ*	*977*	*977*	*977*		*977*	
11	*1.544*	*ℓ*	*1.544*	*1.544*	*1.544*	*1.544*		
14	*1.389*	*ℓ*		*1.389*			*1.389*	
18	*385*	*ℓ*	*385*			*385*	*385*	*385*
19	*1.156*	*ℓ*				*1.156*	*1.156*	*1.156*
20	*911*	*ℓ*	*911*			*911*	*911*	*911*
		{ $R(s_i, T)$ } :=	***23.247***	***22.466***	***21.301***	***21.240***	***21.180***	***20.963***

Tabelul 25. Ierarhia subiectelor de interes

Ierarhizarea autorilor pe subiecte. Fie $c \in C$ un anumit domeniu de cercetare și fie S o submulțime de subiecte (de interes) din domeniul c., $S \subset S(c)$. Se caută o ierarhizare în interiorul mulțimii de autori care au abordat subiectul $s \in S$, $A(s)$, pentru fiecare subiect în parte. Ierarhizarea dorită se realizează prin căutarea automată a tuturor documentelor din colecțiile bibliotecii care tratează subiectele de interes, prin identificarea automată a autorilor acestor documente și prin determinarea indicilor de notorietate-impact pentru fiecare subiect și pentru fiecare autor în subiectul respectiv.

Pentru fiecare document aflat în colecțiile bibliotecii există înregistrate în depozitul de date: identificatorul documentului, d, valoarea funcției de impact, $I(d)$, autorii documentului, $A(d)$ și notorietatea fiecăruia dintre ei, $\mathcal{E}(A)$, publicația în care a apărut documentul, $p(d)$ și notorietatea acesteia, $\varepsilon(p)$, editura publicației, $e(p)$, și notorietatea editurii, $\varepsilon(e)$, precum și subiectele de interes abordate, $S(d)$, $[d] = [d, I(d), A(d), \mathcal{E}(A), p(d), \varepsilon(p), e(p), \varepsilon(e), S(d)]$.

In general, într-un document, d, sunt abordate mai multe subiecte, $S(d)$. În acest context, interesează doar $S(d) \cap S$ adică numai acele subiecte abordate în d care sunt de interes. Se presupune că pentru fiecare subiect $s \in S$ mulțimea documentelor care abordează subiectul s nu este vidă, $D(s) \neq \varnothing$. Pentru a se asigura căutarea într-un set mai bogat de documente sunt luate în considerație atât documentele care tratează subiectele din S, $D(S)$, cât și documentele care le citează $R(D(S))$. Valorile indicilor de notorietate $\varepsilon(x)$ pentru diferitele entități (autori, publicații, edituri), existente în depozitul de date, au fost preluate din diverse liste de notorietate-expert autorizate.

Pentru determinarea indicilor de notorietate-impact $\mathcal{EI}(d, a)$, pentru fiecare document și pentru fiecare autor al documentului respectiv, se utilizează valorile normalizate în intervalul $[0, 1]$ atât ale indicilor de notorietate cât și ale funcțiilor de impact, $\tilde{\varepsilon}(x)$ și respectiv $\tilde{I}(d)$. În

formulele de calcul intervin diverse ponderi: w_a pentru autori, w_p pentru publicații, w_e pentru edituri precum și w_I pentru indicii de impact sau $w_{\mathcal{E}}$ pentru indicii de notorietate. Atribuirea de valori acestor ponderi revine utilizatorului. După determinarea indicilor de notorietate-impact $\mathcal{E}I(d, a)$, pentru fiecare document și pentru fiecare autor al documentului respectiv, în funcție de subiectul ales, s, procedura selectează mulțimea de indici de notorietate-impact, $\mathcal{E}I(s, a)$, aferentă subiectului respectiv. În final, procedura furnizează, în ordinea descrescătoare a indicilor de notorietate-impact, $\mathcal{E}I(s, a)$, lista autorilor care au tratat subiectul s, $a \in A(s)$.

Ierarhia autorilor pentru subiectul de interes ($s = 1$)											
	$a \in A(1)$	$\mathcal{E}I(1, a)$		$a \in A(1)$	$\mathcal{E}I(1, a)$		$a \in A(1)$	$\mathcal{E}I(1, a)$		$a \in A(1)$	$\mathcal{E}I(1, a)$
1	*6*	*0,6656*	**9**	*8*	*0,5686*	**17**	*35*	*0,3279*	**25**	*3*	*0,2599*
2	*27*	*0,6333*	**10**	*41*	*0,5675*	**18**	*22*	*0,3027*	**25**	*43*	*0,2599*
3	*20*	*0,6104*	**11**	*16*	*0,5474*	**19**	*5*	*0,2706*	**27**	*21*	*0,2565*
4	*17*	*0,6051*	**12**	*24*	*0,5229*	**20**	*42*	*0,2704*	**28**	*31*	*0,2478*
5	*10*	*0,6003*	**13**	*36*	*0,4901*	**21**	*46*	*0,2704*	**29**	*19*	*0,2467*
5	*18*	*0,6002*	**14**	*14*	*0,4757*	**22**	*37*	*0,2616*	**30**	*15*	*0,2425*
7	*32*	*0,5805*	**15**	*47*	*0,3584*	**23**	*38*	*0,2616*	**31**	*23*	*0,2403*
8	*34*	*0,5739*	**16**	*39*	*0,3309*	**24**	*2*	*0,2599*			

Tabelul 30 (a). Ierarhie autori pe subiectul *1*.

Ierarhia autorilor pentru subiectul de interes ($s = 5$)											
	$a \in A(5)$	$\mathcal{E}I(5, a)$		$a \in A(5)$	$\mathcal{E}I(5, a)$		$a \in A(5)$	$\mathcal{E}I(5, a)$		$a \in A(5)$	$\mathcal{E}I(5, a)$
1	*16*	*0,7907*	**8**	*34*	*0,5739*	**15**	*2*	*0,2814*	**22**	*31*	*0,2478*
2	*22*	*0,6179*	**9**	*19*	*0,5672*	**16**	*4*	*0,2784*	**23**	*28*	*0,2467*
3	*20*	*0,6104*	**10**	*36*	*0,4901*	**17**	*18*	*0,2716*	**24**	*13*	*0,2446*
4	*17*	*0,6051*	**11**	*47*	*0,3584*	**18**	*5*	*0,2706*	**25**	*8*	*0,2377*
5	*10*	*0,6003*	**12**	*40*	*0,3205*	**19**	*37*	*0,2616*	**26**	*32*	*0,2377*
6	*9*	*0,5953*	**13**	*42*	*0,3205*	**20**	*21*	*0,2565*			
7	*38*	*0,5808*	**14**	*27*	*0,3009*	**21**	*24*	*0,2565*			

Tabelul 30 (b). Ierarhie autori pe subiectul *5*.

Gruparea documentelor după conținut. În general, gruparea se referă la identificarea de grupuri sau clustere într-o mulțime de entități utilizând similarități sau distanțe între acestea.

Fie D un corpus format din k documente. Mulțimea T de termeni (cuvinte sau descriptori) care apar în corpusul de documente, D, formează un vocabular. Fiecare document din corpus, $d \in D$, poate fi considerat ca o listă de termeni din acest vocabular. Presupunând că vocabularul corpusului conține $|T| = n$ termeni, un document oarecare, $d \in D$, poate fi reprezentat printr-un vector n-dimensional, în care fiecare componentă a vectorului este asociată cu un termen din vocabular. Pentru determinarea similarității dintre vectorii n-dimensionali u și v metrica utilizată este similaritatea cosinus care presupune calculul cosinusurilor unghiurilor dintre fiecare pereche de vectori: $cos(\theta_{uv}) = \langle u, v \rangle / \|u\|\|v\|$. În acest scop, fiecărui indice t_i, corespunzător unui termen prezent în d, i se asociază valoarea $f(t_i, d)$ reprezentând frecvența termenului t_i în documentul d: $\varphi(d) = (f(t_1, d), f(t_2, d), \ldots, f(t_n, d))$. Luând în considerare toate perechile de documente din corpusul D rezultă matricea de similaritate cosinus, $M^c \in \mathcal{M}_{k \times k}(\mathcal{R})$, cu elementele: $m^c_{i\ell} = cos(\theta_{i\ell}) = \langle d_i, d_\ell \rangle / \| d_i \| \| d_\ell \|$. Se pot defini și alte matrici de similaritate, utilizând diferiți indici de similaritate: Jaccard, Russel și Rao, etc. Matricile de similaritate asociate celor $k(k - 1)$ perechi de documente pot fi utilizate într-o procedură de grupare bazată pe densitate. Grupurile îndeplinesc condițiile de

omogenitate în raport cu conținutul de termeni. Procedura de grupare bazată pe densitate, derivată din [13], încearcă să identifice și să separe regiunile foarte populate (dense) ale unei mulțimi de puncte, *P*, dintr-un spațiu multidimensional.

Fie r_ε o regiune de căutare de dimensiune ε specificată, numită ε-vecinătate, și fie pr_ε mulțimea punctelor existente în regiunea de căutare. Densitatea este definită de numărul de puncte, nr_ε, din regiunea de căutare, r_ε . Un punct, *B*, este considerat punct de bază dacă, $r_\varepsilon(B)$, ε-vecinătatea sa, conține mai multe puncte decât un număr minim de puncte, p_{min}, specificat de utilizator, $nr_\varepsilon(B) \geq p_{min}$. Punctele de bază sunt în interiorul unui cluster. Un punct , *F*, este punct de frontieră dacă ε-vecinătatea sa $r_\varepsilon(F)$ conține un număr de puncte mai mic decât p_{min}, $nr_\varepsilon(F) < p_{min}$, dar punctul *F* se află în ε-vecinătatea unui punct de bază: $F \in r_\varepsilon(B)$. Un punct, *Z*, este considerat punct de zgomot dacă nu este nici punct de bază și nici punct de frontieră.

Definirea clusterelor se bazează pe noțiunea de accesibilitate în densitate. Un punct *Q* este direct accesibil în densitate dintr-un alt punct *P*, dacă *Q* este conținut în ε-vecinătatea lui *P* și dacă *P* este punct de bază. *P* și *Q* fac parte din același cluster. Un punct *Q* este *accesibil* în densitate dintr-un alt punct *P* dacă există o secvență de puncte $P_1, ..., P_n$ cu $P_1 = P$ și $P_n = Q$ în care fiecare punct P_{i+1} , $i = 1 \div n - 1$, este direct accesibil în densitate din punctul P_i. Relația de accesibilitate în densitate nu este simetrică. Datorită acestei asimetrii, a fost necesară utilizarea noțiunii de conectare în densitate. Două puncte *P* și *Q* sunt conectate în densitate dacă există un punct *O* astfel încât ambele puncte *P* și *Q* sunt accesibile în densitate din *O*. Conectarea în densitate este simetrică. Un cluster este o submulțime de puncte a lui $\mathcal{P}$ care satisface două proprietăți: toate punctele din cluster sunt reciproc conectate în densitate; dacă un punct este conectat în densitate cu orice alt punct din cluster atunci aceasta aparține clusterului.

Entitățile formează un nor de puncte, $P \in \mathcal{P}$, în $\mathcal{R}^n$ înzestrat, în general, cu distanța euclidiană. Distanțele dintre două puncte pot fi determinate fie, direct, utilizând componentele vectorilor $\overline{OP}$, fie utilizând similaritatea cosinus. Pentru vizualizarea rezultatelor este necesar ca norul de puncte, $\mathcal{P}$, să fie situat în $\mathcal{R}^2$ caz care nu reduce din generalitate deoarece, în urma unei analize în componente principale, un nor de puncte din $\mathcal{R}^n$ poate fi proiectat, cu deformări minime, în $\mathcal{R}^2$.

În final, procedura furnizează lista grupurilor identificate, lista punctelor repartizate în fiecare grup, lista punctelor de zgomot și reprezentările grafice ale norului de puncte, înainte și după procesul de grupare.

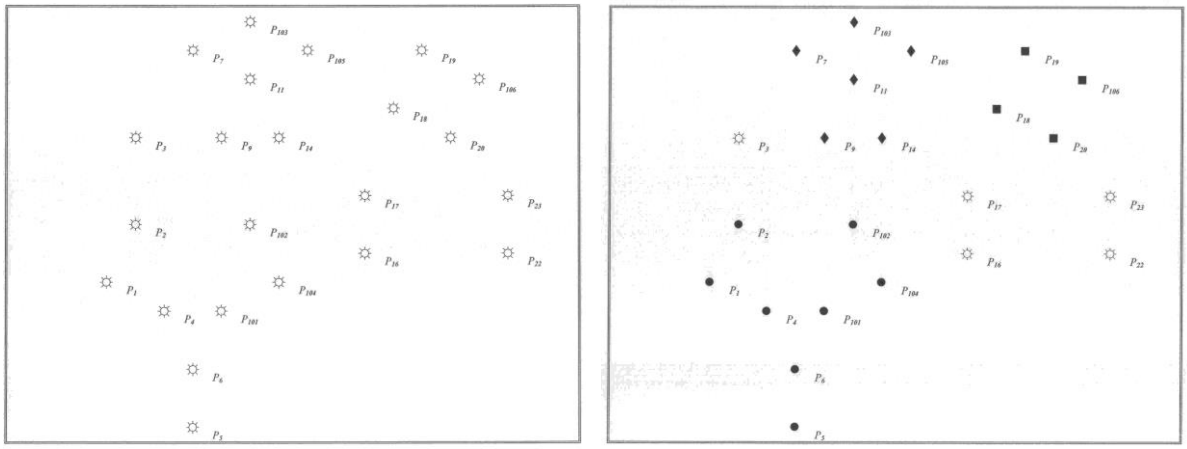


Figura 24. Norul de puncte și grupurile descoperite

Elaborare de recomandări. Avându-se în vedere documentele recent intrate în colecțiile bibliotecii, elaborarea de recomadări către utilizatori constă în identificarea automată a acelor documente care se potrivesc cel mai bine cu interesul fiecărui utilizator în parte. O astfel de identificare este făcută în funcție de comportamentul fiecărui utilizator, respectiv, pe baza operațiilor de consultare documente, pentru fiecare document și fiecare utilizator și de conținutul fiecărui document, respectiv, pe baza listei de descriptori de conținut asociată fiecărui document (termenii din rezumat, cuvintele cheie sau termenii din textul integral).

Fie U mulțimea utilizatorilor activi din perioada de timp analizată T, fie $D^{\ell}(T)$ mulțimea documentelor consultate de aceștia în perioada T și fie $D^{a}(t)$ mulțimea de documente achiziționate (recent) în intervalul de timp t. Pentru fiecare operație (de consultare sau de achiziție, $\theta \in O(\Theta^{\ell}, t)/O(\Theta^{a}, t)$), există înregistrate în depozitul de date: documentul consultat/achiziționat, $d \in D^{\ell}(T)/D^{\ell}(T)$, realizatorul operației (utilizator sau furnizor), u/f, momentul realizării operației, $t \in T$,$[\theta] = [d, u/f, t]$ precum și descriptorii de conținut asociați fiecărui document.

Se presupune că utilitatea, $v(u, d)$, a unui document $d \in D^{\ell}(T)$ pentru un utilizator $u \in U$, este dată de valoarea funcției de realizare a operației θ_{du}, de consultare a documentului d de către utilizatorul u pentru intervalul de timp T, $v(u, d) = R_{du}(\Theta^{\ell}, T) = R_{du}$. Documentele nou intrate, $d \in D^{a}(T)$, sunt recomandate utilizatorilor u pe baza unor utilități estimate, $\tilde{v}(u, d)$, ale fiecărui document pentru fiecare utilizator.

Pentru determinarea acestor estimări se procedează, mai întâi, la o grupare pe conținut a secțiunii de documente analizate, respectiv, a mulțimii $\mathcal{D} = D^{a}(t) \cup D^{\ell}(T)$), rezultatul obținut fiind $\mathcal{D} = \cup_{g \in \mathcal{G}} g$. In continuare, pentru fiecare cluster identificat, $g \in \mathcal{G}$, se estimează utilitatea acestui cluster pentru fiecare utilizator: $\tilde{V}(u, g) = \sum_{d \in g(t)} v(u, d) = \sum_{d \in g(t)} R_{du} = R_{gu}$, unde $g(\ell) = g \cap D^{\ell}(T)$. Se consideră că utilitatea estimată pentru un utilizator u a unui document, nou intrat și situat în clusterul $g \in \mathcal{G}$, este dată de utilitatea estimată a acelui cluster pentru utilizatorul u: $\tilde{v}(u, d) = \tilde{V}(u, g) = R_{gu}$, $(\forall) d \in g(a) = g \cap D^{a}(T)$.

În final, procedura de recomandare oferă fiecărui utilizator o listă cu documentele nou intrate care se situează în aceleași clustere cu documentele consultate de acesta în perioada de timp analizată. În aceste liste, documentele apar în ordinea descrescătoare a utilităților estimate $\tilde{v}(u, d)$.

Gradul de recomandare a documentelor nou intrate pe fiecare dintre utilizatori														
u	*d*	*nota*	*u*	*d*	*nota*	*u*	*d*	*nota*	*u*	*d*	*nota*	*u*	*d*	*nota*
1	*101*	*8,0*	***2***	*103*	*5,8*	***3***	*103*	*6,3*	***4***	*103*	*5,0*	***5***	*101*	*10,0*
1	*102*	*8,0*	***2***	*105*	*5,8*	***3***	*105*	*6,3*	***4***	*105*	*5,0*	***5***	*102*	*10,0*
1	*104*	*8,0*	***2***	*106*	*5,8*	***3***	*101*	*5,8*	***4***	*106*	*4,8*	***5***	*104*	*10,0*
1	*103*	*7,9*	***2***	*101*	*5,6*	***3***	*102*	*5,8*	***4***	*101*	*3,6*	***5***	*103*	*6,1*
1	*105*	*7,9*	***2***	*102*	*5,6*	***3***	*104*	*5,8*	***4***	*102*	*3,6*	***5***	*105*	*6,1*
1	*106*	*4,8*	***2***	*104*	*5,6*	***3***	*106*	*4,7*	***4***	*104*	*3,6*	***5***	*106*	*4,1*

Tabelul 38. Listă de recomandare către utilizatori a noilor documente

CONCLUZII

C1. Concluzii generale

Sistemele suport pentru decizii oferă cunoştinţe şi capacităţi de prelucrare a cunoştinţelor, esenţiale atât în sesizarea situaţiilor decizionale cât și în elaborarea deciziilor și relaxează

limitele cognitive, temporale, spaţiale sau economice ale factorului de decizie. Ele îmbunătăţesc procesele decizionale şi rezultatele luării deciziilor şi se caracterizează prin rolurile pe care le joacă în procesele decizionale.

Un proces decizional: se desfăşoară în etape; conţine un anumit mecanism decizional; poate avea o infrastructură predefinită sau improvizată; poate fi simplu şi stabil sau poate fi un proces adaptiv complex; poate implica atât acţiuni ale unuia sau mai multor sisteme suport pentru decizii cât şi ale uneia sau mai multor persoane fizice (sponsorul, participanţii, implementatorul, alimentatorul şi consumatorul).

Implicarea sistemelor suport pentru decizii în procesele decizionale afectează atât procesele cât şi rezultatele acestora permițându-le să se desfăşoare: cu o productivitate mai ridicată (mai rapid, mai ieftin, cu mai puţin efort); cu o mai mare agilitate (vigilenţă peste aşteptări, mai mare capacitate de răspuns); cu un grad de inovare mai înalt (perspectivă mai clară, creativitate, noutate, surpriză); cu un plus de obiectivitate (precizie mai mare, etică, calitate, încredere) și cu o mai mare cu satisfacţie pentru factorii implicaţi, în comparaţie cu ceea ce s-ar putea obţine dacă nu s-ar recurge la un astfel de suport informatic.

Arhitectura generală a sistemelor suport pentru decizii poate fi descrisă printr-un model conceptual generic care identifică componentele esenţiale ale sistemelor şi interdependenţele acestora. Aceste componente sunt sisteme de diferite tipuri configurate în funcție de specificul fiecărui sistem suport pentru decizii. Între sistemele suport pentru decizii există diferenţieri semnificative determinate de domeniile de aplicabilitate, de caracteristicile de utlizare, de funcţionalităţile proiectate, de abordările privind interacţiunile dintre componente, de modalităţile de încorporare în procesele decizionale, de tipurile de beneficii rezultate din utilizare.

Arhitecturile personalizate păstrează caracteristicile sugerate de cadrul generic dar sunt specializate pe o anumită tehnologie de reprezentare şi prelucrare de cunoştinţe. Dacă factorul decizional are nevoie de capacităţile de prelucrare oferite de mai multe tehnologii de management al cunoştinţelor poate opta pentru utilizarea: fie a mai multor sisteme suport pentru decizii, fiecare orientat către o anumită tehnologie, fie a unui singur sistem suport pentru decizii dar care integrează mai multe tehnologii. Un caz special de integrare, deosebit de important prin implicaţiile sale, rezultă din combinaţia dintre o tehnologie de management a bazelor de date şi o tehnologie de management a rezolvatoarelor flexibile. În acest context, combinarea depozitării datelor cu rezolvatoarele analitice și cu rezolvatoarele *data mining* permite generarea de cunoştinţe noi, deosebit de utile în luarea deciziilor.

Proiectarea conceptuală a depozitelor de date poate fi obținută prin mai multe categorii de metode: orientate către date, orientate către cerinţe și metode mixte sau hibride. Rezultatele cele mai promițătoare au fost obținute prin metodele hibride secvențiale. Etapele generale ale unei astfel de metode sunt: definirea obiectivelor organizației și deducerea cerinţelor informaționale, modelarea multidimensională a datelor, generarea arborilor de atribute (cuburile de date) prin reconciliere cu sursele de date și modelarea avansată a datelor.

Metodele și tehnicile *data mining*, exploratorii şi explicative, reprezintă instrumentele de bază ale prospectorului de date. Produsele informatice comerciale oferă o anumită integrare a acestora (mai mult sau mai puţin completă, mai mult sau mai puţin convivială) în vederea utilizării. Înlănţuirea acestor tehnici trebuie, totuși, făcută conform unei strategii *data mining* care constă, în general, din succesiunea a patru etape: extracţie (extragerea datelor și

asigurarea calităţii acestora); explorare (selecţia, verificarea datelor şi a coerenţei lor, studiul distribuţiilor și relațiilor neliniare, transformări ale variabilelor, selecţionarea acelora cel mai strâns legate de variabila ţintă, completarea datelor care lipsesc); analiză, respectiv, clasificare (caracterizarea claselor prin variabilele iniţiale cu ajutorul instrumentelor de discriminare, nici o variabilă de explicat) și modelare/discriminare (o variabilă de explicat, extracţia unui eşantion de test, estimarea și optimizarea modelelor pentru fiecare din metodele utilizate, compararea performanţelor modelelor optimale, alegerea metodei și a modelului asociat); exploatare (estimarea finală a modelului ales, utilizare curentă şi difuzare rezultate).

Strategia de *data mining* depinde în esenţă de tipurile de variabile considerate şi de obiectivele urmărite. Variabilele considerate sunt de două tipuri, explicative (cantitative, calitative sau mixte, după caz) şi de explicat (nicio variabilă, o variabilă cantitativă Y , binară Z sau calitativă T, după caz). Obiectivele principale urmărite sunt trei: explorarea multidimensională sau reducerea de dimensiune (deducerea unei submulţimi de variabile reprezentative sau a unei mulţimi de componente, prealabile pentru o anumită metodă) și reprezentarea grafică; clasificarea sau segmentarea (deducerea unei variabile calitative); modelarea, Y sau Z sau discriminarea, Z sau T (deducerea unui model de previziune pentru Y, Z sau T). Metodele utilizabile se grupează în funcţie de obiective (explorare, clasificare, modelare), de tipurile variabilelor predictive şi de tipurile variabilelor ţintă.

Performanța unui model, rezultat al unei metode, se evaluează prin capacitatea sa de previziune sau de generalizare. Măsurarea acestei performanțe este foarte importantă deoarece permite să se opereze o selecție de model dintr-o familie asociată metodei utilizate, ghidează alegerea metodei comparând modelele optimale aferente fiecărei metode și oferă o măsură a calității sau a încrederii care se poate acorda previziunii. Estimarea calității previziunii este un element central al oricărei strategii de *data mining*.

Alegerea unui model depinde de mai mulți factori între care complexitatea modelului anvizajat, dimensiunea eșantionului inițial, varianța erorii, complexitatea algoritmilor adică volumul de calcule admisibil. Dacă modelul este cu finalitate explicativă alegerea modelului se bazează pe criterii de ajustare bazate pe ipoteze probabiliste. Dacă obiectivul este esenţialmente predictiv alegerea modelului se bazează pe criterii de calitate a previziunii şi vizează căutarea de modele parcimonioase a căror interpretabilitate trece în plan secundar. Dacă ipotezele, relative atât la modele cât şi la distribuţii, sunt verificate atunci modelele liniare oferă maximum de verosimilitate. Dacă ipotezele distribuţionale nu sunt verificate, dacă relaţiile presupuse între variabile nu sunt liniare sau dacă volumul de date este important atunci devin alternative credibile metode precum reţelele neuronale, maşinile cu support vectorial, cei mai apropiaţi k vecini, arborii de decizie, etc.

Cunoașterea limitelor unui model este extrem de importantă pentru prospectorul de date. În demersul exploratoriu pot fi găsite relaţii între variabile 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. În demersul explicativ, de modelare, o supraparametrizare sau o supraajustare a unui model poate explica perfect datele fără ca rezultatele să fie totuşi extrapolabile sau generalizabile la alte date decât cele studiate. Rezultatele previziunii pot fi viciate de o importantă eroare relativă legată de varianţa estimaţiilor parametrilor, soluția este de a găsi un compromis bun între *bias*-ul unui model mai mult sau mai puţin fals şi varianţa estimatorilor. Trebuie insistat pe fazele, indispensabile, de alegere a metodelor şi de comparare a modelelor optimale.

O bună practică de *data mining* impune asistenților decizionali să cunoască și să știe să articuleze corespunzător toate metodele. Sarcină dificilă, care nu poate fi îndeplinită decât cu condiţia de a avea foarte bine clarificate obiectivele studiului.

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 de variabile auxiliare (artificiale). 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 construcţia modelelor (singurele optimale).

Metodele nu prezintă toate aceleaşi facilităţi de interpretare. Nu există o cea mai bună alegere à priori, numai experienţa şi un protocol de test îngrijit 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 client potenţial şi chiar şi modei.

Obiectivul esenţial rămâne „căutarea sensului” în vederea facilitării luărilor de decizie, prezervând fiabilitatea. Prezenţa sau controlul unei expertize statistice rămâne inevitabilă pentru că necunoaşterea limitelor şi capcanelor metodelor utilizate poate conduce la aberaţii de natură să discrediteze demersul, facând caduce investiţiile consimţite.

Succesul unui proiect, din orice domeniu de activitate al organizaţiilor contemporane, este de multe ori compromis de propensiunea generală de a elabora soluţiile înainte de a identifica şi formula problemele.

Provocările cu care se confruntă un sistem suport pentru decizii de bibliotecă sunt: elaborarea de politici de achiziție orientate către cerere; optimizarea fluxurilor și alocării resurselor; îmbunătățirea conservării colecțiilor; elaborarea de politici de diseminare orientate către cerințe; diseminarea informațiilor/cunoștințelor către utilizatorii potriviți la momentul potrivit; creșterea satisfacției utilizatorilor în sediul propriu și în afara lui; diversificarea și creșterea veniturilor culturale și comerciale; comunicarea mai bună cu partenerii.

Domeniile de activitate ale bibliotecii, care pot fi îmbunătățite, sunt: dezvoltarea colecţiilor; accesibilitatea colecțiilor; accesul la publicații; utilizarea bibliotecii; digitalizarea colecțiilor; serviciile bibliografice, asistența și îndrumarea; potențialul de dezvoltare; conservarea colecţiilor; managementul.

Utilizatorii sistemului solicită un spectru larg de expertize, de la căutări simple la statistici avansate. Pentru a putea adapta seviciile oferite de sistem la cerințele fiecărei categorii de utilizatori principalele categorii de servicii care trebuiesc avute în vedere sunt: căutările simple, care furnizează rapoarte predefinite și valori ale indicatorilor operaționali și/sau de performanță; interogările avansate și/sau personalizate; analizele avansate, care implică navigare multidimensională și funcții puternice de analiză; simulările și statisticile avansate.

Arhitectura necesară sistemului suport pentru decizii al unei biblioteci este combinația dintre tehnologia de management a bazelor de date şi tehnologia de management a rezolvatoarelor flexibile capabilă să asigure integrarea depozitării datelor cu rezolvatoarele analitice şi rezolvatoarele *data mining*.

Pentru realizarea depozitului de date este foarte important ca proiectanții să urmeze o metodologie de proiectare conceptuală consolidată şi robustă dat fiind că dezvoltarea acestuia este un proces foarte scump chiar în condițiile actuale când există instrumente software care oferă soluţii prefabricate acoperind toate etapele din ciclul de viaţă al unui depozit de date.

Pentru implementarea aplicațiilor de *data mining* trebuie urmată o strategie simplă și eficientă pentru definirea obiectivelor, selecția variabilelor semnificative, alegerea metodelor și modelelor asociate, asigurarea calității datelor utilizabile, estimarea calității și fiabilității rezultatelor.

Concepția și implementarea sistemului suport pentru decizii al unei biblioteci, ca de altfel ale oricărui sistem informatic, sunt influențate de către o serie de factori, între care pot fi menționați: obiectivele urmărite; recomandările, normele și standardele utilizate; restricțiile impuse de către instituție; evoluția mediului; personalul implicat; bugetul disponibil pentru realizare; termenele de finalizare.

Obiectivele sistemului suport pentru decizii de bibliotecă sunt: furnizarea de indicatori care să permită evaluarea în timp a conformității cu obiectivele bibliotecii (evaluarea rezultatelor obținute, sesizarea tendințelor, alerte, evaluarea indicatorilor operaționali, de performanță și bibliometrici, rapoarte de activitate); furnizarea unor instrumente de analiză a tendințelor, de sesizare a situațiilor decizionale și de sugerare a unor acțiuni corespunzătoare (analize complexe, simulări, prognoze) în vederea luărilor de decizii; integrarea datelor și compararea informațiilor din aplicații informatice existente; simplificarea accesului la informație prin schimb transparent de informații și diseminare accelerată a informațiilor.

Pentru stabilirea cerințelor informaționale se impune aplicarea cu discernământ a prevederilor normative specifice domeniului bibliotecilor elaborate, recomandate și utilizate atât pe plan intern cât și pe plan internațional (descrierile bibliografice, indicatorii operaționali, indicatorii de performanță și indicatorii bibliometrici) și definirea unui sistem formalizat, unitar, coerent și evolutiv de indicatori.

Indicatorii bibliometrici se bazează pe ipoteza că frecvența citărilor unui articol de către alte articole reflectă calitatea acelui articol și oferă doar o imagine parțială și părtinitoare a anumitor aspecte ale vieții științifice, fără acoperirea ansamblului. Aceștia trebuie să fie completați și/sau corectați de experții din domeniul științei și, de asemenea, interpretați dacă sunt utilizați în scopul unei evaluări sau luări de decizii. Indicatorii numerici sunt foarte ușor manipulabili de către persoane fizice, instituții și alte părți interesate din viața științifică. Numărul manipulărilor crește și el poate fi corelat cu efectul influenței crescânde a indicatorilor. Utilizarea indicatorilor bazați pe analiza citărilor nu este favorabilă asumării de riscuri științifice și inovării. O utilizare abuzivă a acestora sau, mai rău, automată ar fi un obstacol major în calea inovării.

Pentru evaluarea resurselor financiare necesare construirii sistemului suport pentru decizii aferent unei biblioteci hibride trebuiesc avute în vedere următoarele produse și servicii: instrumentele de proiectare a depozitului de date; instrumentele de extragere, transformare și încărcare a datelor; instrumentele de interogare și raportare; instrumentele de fundamentare a deciziilor (prelucrările analitice, explorarea datelor și descoperirea cunoștințelor din date); contractele de dezvoltare a sistemului și consultanță.

Pentru resursele umane implicate trebuie să se asigure o anumită structură: comitet de management, echipă de proiect, grupuri de lucru cu utilizatorii, grupă de consultanți, un subcontractant, firmă de specialitate, pentru dezvoltare și implementare.

Pentru bibliotecă avantajele majore ale sistemului suport pentru decizii sunt: asigură informații de calitate și noi instrumente de management; rezolvă faze tehnice critice privind modelarea, furnizarea și stocarea datelor; satisface cerințe tehnice actuale și viitoare; satisface cerințele utilizatorilor; este adaptabil; susține trecerea la o cultură orientată către performanță și impune personalului dezvoltarea în consecință a abilităților; promite mari îmbunătățiri în modul de înțelegere a ceea ce se face în prezent și a ceea ce se preconizează pentru viitor.

Pentru companii avantajele sistemului suport pentru decizii al bibliotecii constau în asigurarea unor puternice funcționalități de alimentator de cunoștințe pentru sistemele suport pentru decizii ale acestora prin diseminarea informațiilor/cunoștințelor către utilizatorii interesați la momentele oportune.

Pentru cercetători și practicieni, care abordează dezvoltarea de sisteme suport pentru decizii pentru diverse companii, modul de abordare și construire a sistemului suport pentru decizii pentru biblioteci oferă un cadru conceptual și metodologic de integrare a depozitării datelor cu analiticile *on-line* și *data mining* care se poate dovedi foarte util în demersurile lor.

Pentru a se oferi şanse cât mai favorabile de succes utilizării tehnologiei *data mining* în sistemele suport pentru decizii este necesar ca preocupările legate de definirea obiectivelor şi de analiză a datelor să intervină cât mai devreme posibil în procesul de construire al oricărui sistem suport pentru decizii. În cazul sistemului decizional al bibliotecii, faptul că cerințele informaționale au fost deduse din setul complet de obiective instituționale și reconciliate cu sursele de date a condus la obținerea, pentru toate procedurile definite, a unor avantaje substanțiale precum: disponibilitatea datelor necesare în depozitul de date, simplificarea consistentă a algoritmilor de calcul sau, mai ales, posibilitatea de a se profita direct de performanțele sistemelor *OLAP*.

În prezenta teză de doctorat s-au adus contribuții semnificative privind susținerea proceselor decizionale, atât dintro organizație de tip bibliotecă hibridă, oferindu-se o integrare eficientă a tehnologiilor *OLAP* și *DMKD* prin intermediul unui singur depozit de date, cât și din alte organizații, conferindu-se sistemului suport pentru decizii al bibliotecii, prin construcție, rolul de principal alimentator de cunoștințe pentru sistemele decizionale ale acestora.

C2. Contribuții

În cadrul tezei s-au adus o serie de contribuții personale ale autorului constând în:

- Definirea menirii/rolurilor sistemului suport pentru decizii, în strictă concordanță atât cu obiectivele instituționale ale bibliotecilor hibride cât și cu principalii factori care influențează concepția și implementarea sistemelor informatice.

- Definirea arhitecturii sistemului suport pentru decizii de bibliotecă, compatibilă cu arhitectura generică a sistemelor suport pentru decizii, bazată integrarea tehnologiei de management a bazelor de date cu tehnologii de management a rezolvatoarelor flexibile (analitice *on-line* și *data mining*) prin un singur depozit de date.

- Definirea și descrierea entităților bibliografice și a relațiilor dintre ele, în concordanță atât cu cerințele funcționale privind datele bibliografice cât și cu modelul relațional al bazelor de date.

- Definirea surogatelor bibliografice pe baza reconcilierii cerințelor informaționale, deduse din obiectivele instituționale ale bibliotecii, cu sursele de date.

- Abordarea formalizată a aspectelor specifice şi măsurabile ale faptelor, relevante pentru analiză, la nivelul minim de granularitate.

- Formalizarea și unificarea sistemului de indicatori (de stare, de performanță și bibliometrici) prin definirea de formule evaluabile pentru toți indicatorii uzuali în strictă concordanță semnificațiile curente ale acestor indicatori.

- Analiza și identificarea elementelor multidimensionale, definirea schemei dimensionale a depozitului de date, modelarea multidimensională a datelor și proiectarea conceptuală a depozitului de date.

- Elaborarea și experimentarea de proceduri de descoperire a cunoștințelor: ierarhizarea preferințelor de lectură ale utilizatorilor, ierarhizarea subiectelor de interes, ierarhizarea autorilor pe subiecte, regăsirea/gruparea documentelor după conținut, recomandarea documentelor către utilizatori.

- Evidențierea principalelor avantaje ale sistemului suport pentru decizii al bibliotecii pentru mediul decizional al instituției, pentru sistemele decizionale ale altor companii precum și pentru cercetători și practicieni care abordează dezvoltarea de sisteme suport pentru decizii pentru diverse organizații.

- Analiza, selecția și sinteza în viziune proprie, subordonată strict obiectivelor cercetării, a materialelor consultate referitoare la sistemele informatice din clasa sistemelor suport pentru decizii, la utilizarea eficientă a tehnologiei de explorare a datelor și descoperire a cunoștințelor în susținerea proceselor decizionale precum și la modelarea multidimensională a bazelor de date.

- Elaborarea și aplicarea unei strategii simple dar eficace de implementare a aplicațiilor de *data mining*.

- Selecția celor mai frecvent utilizate metode/modele de *data mining* și descrierea sintetică a acestora conform strategiei elaborate cu evidențierea aspectelor relevante pentru prospectorul de date.

- Adaptarea și aplicarea unei metode consolidate și robuste de proiectare conceptuală a depozitelor de date bazată pe abordarea hibridă secvenţială.

- Adoptarea și rafinarea unei soluții conceptuale optimale de modelare multidimensională a bazelor de date, neifluențată de contextele și aspectele particulare de implementare.

- Abordarea sistemică a mediului decizional și a situaţiilor decizionale, cu focalizare pe deciziile manageriale și adoptarea deciziilor prin metode ştiinţifice, dintro perspectivă modernă și cu un grad ridicat de conceptualizare şi de generalitate.

C.3 Direcții viitoare ale cercetării

Volumul datelor stocate astăzi este în plină expansiune, datele numerice create în lume au evoluat de la *1,2* zettaocteți de date în anul *2010*, la *1,8* zettaocteți în *2011*, apoi la *2,8* zettaocteți în *2012* și se estimează la *40* zettaocteți în *2020* (*1 Zo = 10^{21}*).

Fenomenul *big data*, respectiv, aceste noi ordine de mărime, evidențiază necesitatea ca preluarea, stocarea, cercetarea, partajarea, analiza și vizualizarea datelor să fie regândite și redefinite. Experții au considerat fenomenul *big data* drept o provocare informatică majoră a deceniului *2010-2020* și o nouă prioritate a cercetării și dezvoltării.

Managementul datelor devine un proces foarte complex datorită faptului că volume imense de date provin din surse multiple. Se impune ca aceste date să fie relaționate, conectate și corelate, pentru ca procesul să fie capabil „să perceapă" informația care se presupune că este transmisă prin aceste date.

Tehnologia *big data* procesează și analizează aceste date la volumul și viteza dorită. Scopul tehnologiei *big data* este să analizeze toate datele disponibile, eficient din punct de vedere costuri. Orice date, așa cum sunt. Se pot analiza date structurate, video, audio, date spațiale sau orice tip de date.

Noile modele de reprezentare a datelor (*Big Data Architecture framework - BDAF*) permit garantarea performanțelor pentru volumele de date în cauză. Au fost propuse structuri bazate pe servere standard ale căror configurații sunt optimizate. Cererile sunt descompuse, distribuite nodurilor paralelizate, executate în paralel iar rezultatele reunite și recuperate. Cercetările se concentrează către sisteme cu o puternică scalabilitate orizontală și către soluții bazate pe *NoSQL*.

Pentru a răspunde problematicii *big data* arhitectura de stocare a sistemelor este regândită și modelele de stocare se multiplică în consecință: *Cloud computing, High Performance Computing, Distributed files system.*

Maturizarea subiectului a condus la evidențierea unui criteriu, mult mai profund, de diferențiere dintre informatica decizională și *big data* în ceeace privește datele și utilizarea acestora:

- Informatica decizională utilizează statistica descriptivă, pentru date cu mare densitate în informație, pentru a măsura fenomene, a detecta tendințe, etc;
- *Big data* utilizează statistica inferențială, pentru date cu slabă densitate în informație, ale căror volume, foarte mari, permit inferențe ale legilor (regresii, ...) conferindu-le capacități predictive (cu limitele acestor inferențe).

În concluzie, se conturează două direcții distincte de continuare a cercetărilor:

- în plan teoretic, direcția de contiuare firească a cercetărilor o constituie problematica asistării deciziilor bazată pe *big data mining*;
- în plan practic, trecerea la experimentarea și implementarea etapizată a sistemului suport pentru decizii al bibliotecii pe baze contractuale datorită resurselor umane și financiare implicate de un astfel de demers.

BIBLIOGRAFIE SELECTIVĂ

9 BERSON, Alex; SMITH, Stephen J. *Building data mining applications for CRM*. McGraw-Hill, Inc., 2002.

10 BESSE, Philippe. *Exploration Statistique Multidimensionnelle*. Institut National des Sciences Appliquées de Toulouse, 2014.

11 BESSE, Philippe; LAURENT, Beatrice. *Apprentissage Statistique*: *modélisation, prévision et data mining*. Institut National des Sciences Appliquées de Toulouse, 2014.

14 BORNE, Pierre; et al. *Optimisation en sciences de l'ingénieur: Méthodes exactes*. Lavoisier, 2013.

16 BURSTEIN, Frada; HOLSAPPLE, Clyde (ed.). *Handbook on decision support systems, 1: Basic Themes, 2: Variations*. Springer Berlin Heidelberg, 2008.

21 CHU, Wesley W. (Ed.). *Data Mining and Knowledge Discovery for Big Data*: *Methodologies, Challenge and Opportunities*. Springer Berlin Heidelberg, 2014.

29 DIMA, Ioan Constantin; MAN, Mariana. *Modelling and Simulation in Management: Econometric Models Used in the Management of Organizations*. Springer, 2015.

35 ENĂCHESCU, Denis. *Data Mining: metode şi aplicaţii*. Editura Academiei, 2009.

39 FILIP, Florin Gheorghe. *Decizie asistată de calculator: decizii, decidenți, metode și instrumente de bază*. Editura Tehnică, București, 2002.

40 FILIP, Florin Gheorghe. *Sisteme suport pentru decizii*. Ed. 2, Editura Tehnică, București, 2007.

50 GOLFARELLI, Matteo; RIZZI, Stefano. *Data Warehouse design: Modern principles and methodologies*. McGraw-Hill, Inc., 2009.

53 GORUNESCU, Florin. *Data Mining: Concepts, models and techniques*. Springer Science & Business Media, 2011.

54 HAN, Jiawei; KAMBER, Micheline; PEI, Jian. *Data mining: concepts and techniques*. Elsevier, 2011.

55 HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. *The elements of statistical learning: data mining, inference, and prediction*. 2nd Ed., Springer, 2009.

61 HOLSAPPLE, Clyde; WHINSTON, Andrew B. (ed.). *Recent developments in decision support systems*. Springer Science & Business Media, 2013.

64 INMON, William H. *Building the data warehouse*. John wiley & sons, 2005.

79 LEE, Keun-Woo; HUH, Soon-Young. *Model-solver integration in decision support systems: a web services approach*, 2003.

83 LEPĂDATU, Cornel. Support Systems for Knowledge Culture based on Solution and Tools from the Field of Business Intelligence - SSCBI. In *Proceedings of the Workshop IST – Multidisciplinary Approaches*, Bucharest, Romania, 2006 : 7-12.

84 LEPĂDATU, Cornel. Acquisition Policy of a Library and Data Mining Techniques. *Studies in informatics and control*, 2007, 16(4) : 413-420.

87 LEPĂDATU, Cornel. Explorarea datelor şi descoperirea cunoştinţelor - probleme, obiective şi strategii. *Revista Română de Informatică şi Automatică*, 2012, 22.4 : 5-14.

88 LEPĂDATU, Cornel. Metode exploratorii multidimensionale. *Revista Română de Informatică şi Automatică*, 2013, 23.1 : 14-30.

90 LEPĂDATU, Cornel. Sisteme suport pentru decizii şi bibliomining. *Revista Română de Informatică şi Automatică*, 2014, 24.2 : 17-30.

91 LEPĂDATU, Cornel. Sistem suport pentru decizii de bibliotecă. *Revista Română de Informatică şi Automatică*, 2014, 24.3 : 5-17.

92 LEPĂDATU, Cornel. Descoperirea cunoștințelor din date: metode predictive. *Revista Română de Informatică şi Automatică*, 2015, 25.3 : 57-74.

101 MAIMON, Oded; ROKACH, Lior (ed.). *Data mining and knowledge discovery handbook*. New York,

Dordrecht, Heidelberg, London: 2nd Ed., Springer, 2010.

117 MINER, Gary; NISBET, Robert; ELDER IV, John. *Handbook of statistical analysis and data mining applications*. Academic Press, 2009.

132 PENG, Yi, et al. A descriptive framework for the field of data mining and knowledge discovery. *International Journal of Information Technology & Decision Making*, 2008, 7.04: 639-682.

172 PHILIP, S. Yu; HAN, Jiawei; FALOUTSOS, Christos. *Link Mining: Models, Algorithms, and Applications*. Springer, 2010.

134 POWER, Daniel J. *Decision Support Systems: Concepts and Resources for Managers*. NY: Greenwood Publishing Group, 2002.

135 POWER, Daniel J. *Decision support, analytics, and business intelligence*. Business Expert Press, 2013.

140 RAFANELLI, Maurizio (ed.). *Multidimensional Databases: Problems and Solutions*. Idea Group Inc., 2003.

143 RICCI, Francesco; ROKACH, Lior; SHAPIRA, Bracha. *Introduction to recommender systems handbook*. Springer US, 2011.

153 SPRAGUE JR, Ralph H.; CARLSON, Eric D. *Building effective decision support systems*. Prentice Hall Professional Technical Reference, 1982.

154 SPRAGUE JR, Ralph H.; WATSON, Hugh J. *Decision Support Systems: Putting Theory into Practice*. 3rd edition, Prentice Hall, 1993.

158 TUFFÉRY, Stéphane. *Data mining et statistique décisionnelle: l'intelligence des données*. 4ème edition, Editions Technip, 2012.

159 TUFFÉRY, Stéphane. *Modélisation Predictive et Apprentissage Statistique avec R*. Editions Technip, 2015.

160 TURBAN, Efraim; MEREDITH, Jack R. *Fundamentals of management science*. McGraw-Hill College, 1998.

161 TURBAN, Efraim; SHARDA, Ramesh; DELEN, Dursun. *Decision support and business intelligence systems*. 9th Edition, Prentice-Hall Inc., 2011.

162 VAISMAN, Alejandro; ZIMÁNYI, Esteban. *Data Warehouse Systems: Design and Implementation*. Springer, 2014.

164 VAPNIK, Vladimir Naumovich. *Statistical learning theory*. New York: Wiley, 1998.

171 WU, Xindong; KUMAR, Vipin (ed.). *The top ten algorithms in data mining*. CRC Press, 2009.