DESCOPERIREA CUNOŞTINŢELOR DIN DATE: METODE PREDICTIVE

Cornel Lepădatu

cornel_lepadatu@biblacad.ro

Biblioteca Academiei Române - Bucureşti

Rezumat. Obiectivul principal al metodelor predictive îl constituie căutarea de modele optimale pentru diferite metode de modelare: clasice (regresia multiplă, analiza discriminantă), mai puțin clasice (segmentarea) sau de instruire (rețelele neuronale, agregarea de modele, maşinile cu suport vectorial). Articolul se concentrează pe prezentarea sub o formă omogenă şi sintetică a celor mai frecvent utilizate metode de instruire supervizată pentru descoperirea de cunoştinţe din volume (foarte) mari de date (*Big data*, *DCD*) pentru sprijinirea deciziilor în diverse domenii de aplicare. Pentru fiecare metodă au fost evidențiate, după caz, o serie de aspecte specifice esențiale pentru prospectorul de date: domeniile de aplicabilitate, semnificațiile coeficienților, puterea de discriminare a caracteristicilor, metodele de selecție a variabilelor, adecvarea modelului cu datele observate, măsurarea performanțelor, separarea estimării modelului de estimarea erorilor de previziune, controlul supra-învățării, caracterizarea şi interpretarea rezultatelor, performanțele computaționale.

Cuvinte cheie: big data, descoperire cunoştinţe din date (*DCD*), discriminare, instruire, modelare, previziune.

Abstract: The main objective of predictive methods is the search for optimal models for various modeling techniques: classical (multiple regression, discriminant analysis), less classic (classification and regression tree) or machine learning (neural networks, ensemble methods, support vector machines). The article focuses on an uniform and synthetic presentation for the supervised learning methods the most commonly used for knowledge discovery from (very) large amount of data (*Big data, KDD*) for decision support in various fields of application. For each method were highlighted, as appropriate, a number of specific issues, essential for an data prospector: the fields of the application, the significances of the coefficients , the discrimination power of the characteristics, the methods for selection of variables, the appropriateness of the model with the observed data, the performances measurement, the separation of model estimation error from the prediction estimation errors, over-learning control, characterization and interpretation of results, computational performances.

Keywords: Big data, Classification, Knowledge Discovery in Databases (*KDD*), Modeling, Prediction, Statistical learning.

1. Introducere

Mediul economic, social şi politic în care se iau în prezent deciziile se caracterizează printr-o dinamică pronunțată şi continuă în care tehnologiile avansate devin un determinant major al stilului de viață uman. Numărul căilor de acțiune posibile poate fi foarte mare, gradul de incertitudine poate face foarte dificilă previziunea consecințelor luării unei decizii, efectele unor erori în luarea deciziilor ar putea fi dezastruoase datorită complexității operațiilor şi reacțiilor în lanț pe care aceste erori pot să le cauzeze [9, 10].

Convergența procesării informației cu tehnicile de comunicații, ilustrată elocvent mai ales prin dezvoltarea exponențială a Internet-ului, a determinat apariția unor enorme cantități de date, informații 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 apariția agresivă a unor tehnologii emergente precum sistemele dedicate (*embeded*), sistemele mobile şi respectiv sistemele omniprezente (*ubiquitous*) de prelucrare a informației [1, 2, 3, 5, 6, 7, 14, 15].

Este, deci, indiscutabil de clară necesitatea extragerii de informații şi de cunoştinţe, din aceste masive de date distribuite, în primul rând pentru asistarea proceselor decizionale. In acest sens, esențial este faptul că este nevoie de a reprezenta în mod explicit caracteristici importante ale informațiilor, care nu mai sunt legate de reprezentarea abstractă a conceptelor lumii reale ci, mai degrabă, de obiectivul factorilor de decizie şi anume susținerea proceselor de analiză a datelor orientate către luarea deciziilor [9, 10, 12].

Menirea sistemelor suport pentru decizii este de a atenua efectul limitelor şi restricțiilor factorului decizional în rezolvarea problemelor decizionale. În desfăşurarea proceselor decizionale poziția centrală este ocupată de intuiția şi judecata umană iar metodele utilizate se bazează pe

analiza datelor disponibile [4, 9, 10, 12, 16, 20, 21, 23].

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*) şi depozitarea datelor (*data warehousing*) precum şi din explorarea datelor şi descoperirea cunoştinţelor (*data mining and knowledge discovery*) [10, 16, 17, 20, 21, 23].

2. Modelare în vederea previziunii

Descoperirea cunoştinţelor din baze de date (*DCD*) vizează căutarea de informaţii relevante pentru previziuni, în vederea susţinerii proceselor decizionale, recurgând la metode de instruire (*machine and statistical learning*) care iau în considerare specificitatea volumelor de date, mari şi foarte mari.

Având la dispoziţie observaţii asupra unei variabile explicative multidimensionale, $\boldsymbol{X} = \{X^j\}_{j=1}^p$, măsurată pe o mulţime de n indivizi, în funcţie de absenţa sau prezenţa unei variabile de explicat Y observată în conjuncţie cu $\boldsymbol{X}$, se pot distinge două tipuri de probleme, numite de instruire:

- probleme de instruire nesupervizată, absenţa variabilei de explicat: să se găsească o tipologie sau taxinomie a observaţiilor, cum să fie acestea grupate în clase cât mai omogene dar cât mai diferite între ele;
- probleme de instruire supervizată sau de modelare, prezenţa variabilei de explicat: să se găsească, observându-l pe $\boldsymbol{X}$, o funcţie φ susceptibilă să reproducă „cel mai bine” (conform unui criteriu ce urmează a fi definit) pe Y, $Y = \varphi(\boldsymbol{X}) + \varepsilon$, unde ε simbolizează eroarea sau zgomotul de măsurare.

Modelarea sau discriminarea, respectiv, deducerea unui model de previziune pentru o variabilă ţintă, constituie obiectivul principal urmărit în demersul inferenţial şi confirmatoriu [8, 18, 19] pentru aplicaţiile *DCD*. Tehnicile cele mai frecvent utilizate în atingerea acestui obiectiv au fost: modelele liniare, analiza discriminantă, reţelele neuronale, maşinile cu suport vectorial, arborii de clasificare şi de regresie, agregarea modelelor [2, 8, 12, 13, 22, 24].

Principalele aspecte evolutive privind modelarea în vederea previziunii au fost [2]:

- anii 1940-1970. Statistică: o problemă, asociată cu o ipoteză discutabilă, un experiment planificat cu $p \leq 10$ variabile observate pe $n \approx 30$ indivizi, un model liniar presupus real, un test, o decizie, un răspuns, un volum de date de ordinul hectooctetilor.
- anii 1970. Se generalizează primele instrumente *IT* : analiza datelor (*multivariate statistics*) explorează, fără a pretinde un model, volume mai mari de date kiloocteţi.
- anii 1980. În inteligenţa artificială sistemele expert, considerate depăşite, sunt înlocuite de instruirea reţelelor neuronale (*ANN learning*) iar în statistică sunt abordate modelele neparametrice sau funcţionale, volumul de date megaocteţi.
- anii 1990. Prima schimbare de paradigmă: datele nu mai sunt planificate, ele sunt culese şi stocate pentru activităţile curente ale companiilor dar sunt valorificate pentru sprijinirea proceselor decizionale [9], este momentul apariţiei marketing-ului cantitativ şi al gestionării relaţiilor cu clienţii (*CRM*), volumul datelor gigaocteţi.
- anii 2000. A doua schimbare de paradigmă: numărul p al variabilelor „explodează”($p >> n$, $p \approx 105$), obiectivul privind calitatea previziunii prevalează asupra realităţii modelului (acesta devenind „cutie neagră”), *machine learning* şi *statistics* se reunesc în *statistical learning* [13, 25], modelele „bune” sunt acelea care asigură un „echilibru” între bias şi varianţă, respectiv, o minimizare coordonată a erorilor de aproximare (bias) şi a erorii de estimare (varianţă), volumul datelor teraocteţi.
- anii 2010. A treia schimbare de paradigmă: în multe aplicaţii numărul n, de indivizi, „explodează”, bazele de date debordează (*big data*) şi sunt structurate în nori (*cloud*),

mediile computaționale sunt grupate (*cluster*) dar puterea acestora tot nu este suficientă pentru complexitatea (*greed*) algoritmilor, apare o a treia noțiune privind eroarea, o optimizare indusă prin limitarea timpului de calcul sau a volumului/fluxului de date luate în considerare, decizia devine adaptivă sau secvențială, volumul datelor este de ordinul petaocteților.

- anul 2014. Comisia Europeană programează două domenii de abordare pentru *big data*: îmbunătățirea capabilității companiilor de a realiza produse şi servicii de date, inovatoare şi multilingve şi rezolvarea problemelor de cercetare (fundamentală şi aplicativă cu orientare către piață) privind capabilitățile analiticilor predictive de a susține scalabilitatea şi capacitatea de reacție [11].

3. Estimarea calității 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 şi oferă o măsură a calității sau a încrederii care se poate acorda previziunii.

Estimarea calității previziunii este un element central al oricărei strategii *DCD*. În principiu, sunt avute în vedere trei tipuri de abordări: partiționarea eşantionului 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 uneia dintre abordări depinde de mai mulți factori între care dimensiunea eşantionului inițial, complexitatea modelului anvizajat, varianța erorii, complexitatea algoritmilor adică volumul de calcule admisibil.

Fie $Y = \varphi(\boldsymbol{X}) + \varepsilon$ modelul de estimat cu ε independent de $\boldsymbol{X}$, $M(\varepsilon) = 0$ şi $var(\varepsilon) = \sigma^2$, fie F legea lui Y în conjuncție cu $\boldsymbol{X}$, şi fie $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}^{\boldsymbol{n}}_{i=1}$ un eşantion. Eroarea de previziune a modelului este definită prin: $\mathcal{E}_P(\boldsymbol{z}, F) = M_F\,[Q(Y, \tilde{\varphi}(\boldsymbol{X}))]$, 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 indice de misclasare $Q(y, \tilde{y}) = 1_{\{y \neq \tilde{y}\}}$.

În cazul cantitativ eroarea de previziune, într-un punct $\boldsymbol{x}$, se descompune astfel:

$\mathcal{E}_P(\boldsymbol{x}) = \sigma^2 + \text{bias}^2 + \text{varianță}$.

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 şi deci bias-ul său va putea fi cu atât mai redus. Dar, pe de altă parte, varianța creşte 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 şi 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 eşantionul observat, este $\mathcal{E}_P = (1/n)\sum^{n}_{i=1} Q(y_i, \tilde{\varphi}(\boldsymbol{x}_i))$.

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 eşantion independent care nu a participat la estimarea modelului.

Dacă dimensiunea eşantionului este suficient de mare, se procedează la separarea eşantionului în trei părți numite de instruire, de validare şi de test $\boldsymbol{z} = \boldsymbol{z}_{ins} \cup \boldsymbol{z}_{val} \cup \boldsymbol{z}_{test}$):

- $\tilde{\mathcal{E}}_P(\boldsymbol{z}_{ins})$ este minimizată pentru a estima modelul;
- $\tilde{\mathcal{E}}_P(\boldsymbol{z}_{val})$ serveşte la compararea modelelor în interiorul unei aceleiaşi familii pentru a-l

selecționa pe acela care minimizează această eroare;

- $\tilde{\varepsilon}_P(z_{test})$ este utilizată pentru a compara între ele cele mai „bune" modele ale fiecărei metode considerate.

Dacă dimensiunea eşantionului este insuficientă calitatea ajustării este degradată, varianța estimării erorii poate fi importantă dar nu poate fi estimată şi atunci selecția modelului se bazează pe un alt tip de estimare a erorii de previziune recurgându-se la simulare (validare încrucişată) sau la penalizare.

În situațiile în care legea F a eşantionului nu este cunoscută sau, de cele mai multe ori, atunci când nu se poate presupune că este gaussiană, evaluarea distribuției estimatorului se face prin simulare recurgându-se la tehnici de bootstrap (sau reeşantionare). 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 aferentă eşantionului de învățare, necunoscută, cu distribuția empirică $\tilde{F}$ obținută acordând o pondere de $1/n$ fiecărei realizări. Astfel se obține un eşantion de dimensiune n, numit eşantion bootstrap, cu legea de distribuție empirică $\tilde{F}$ prin n extrageri aleatoare cu înlocuire dintre cele n observații inițiale. Se poate obține, fără dificultate, un număr mare de eşantioane 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 ale riscului pătratic) şi chiar intervalele de încredere ale estimatorului, fără vre-o ipoteză (normalitate) privind legea reală.

Fie $z^* = \{(x^*_i, y^*_i)\}^n_{i=1}$ un eşantion bootstrap al datelor:

- estimatorul plug-in al erorii de previziune, în care distribuția F este înlocuită cu distribuția empirică $\tilde{F}$, este definit prin: $\varepsilon_P(z^*, \tilde{F}) = (1/n)\sum^n_{i=1} nQ(y^*_i, \varphi z^*(x^*_i))$, unde φz^* reprezintă estimarea lui φ pe z^*;
- estimarea bootstrap a erorii medii de previziune este dată de: $\varepsilon_{boot} = M_{\tilde{F}}[\varepsilon_P(z^*, \tilde{F})] = M_{\tilde{F}}[(1/n)\sum^n_{i=1} nQ(y^*_i, \varphi z^*(x^*_i))]$;
- estimarea obținută prin simulare va fi: $\tilde{\varepsilon}_{boot} = (1/\kappa)\sum^{\kappa}_{k=1}(1/n)\sum^n_{i=1} nQ(y^*_i, \varphi z^*_k(x^*_i))$.

Estimarea erorii de previziune astfel construită este, în general, biasată prin optimism deoarece, datorită simulărilor, aceleaşi observații apar în acelaşi timp şi în estimarea modelului şi în estimarea erorii. Există abordări care vizează corecția acestui bias. Estimatorul out-of-bag al erorii de previziune $\tilde{\varepsilon}_{oob}$, inspirat din validarea încrucişată, consideră, pe de o parte, observațiile extrase în eşantionul bootstrap şi, pe de altă parte, observațiile neutilizate la estimarea modelului dar reținute pentru estimarea erorii:

$$\tilde{\varepsilon}_{oob} = (1/n)\sum^n_{i=1} 1/B_i \sum_{k \in K_i} Q(y^*_i, \varphi z^*_k(x^*_i))$$

unde K_i reprezintă mulțimea de indici k ai eşantioanelor bootstrap neconținând a i-a observație după cele B simulări şi B_i reprezintă numărul $|K_i|$ al acestor eşantioane. B trebuie să fie suficient de mare pentru ca orice observație să poată să fie extrasă cel puțin o dată, altfel termenii cu $K_i = 0$ trebuiesc omişi.

4. 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 varianță) iar dacă mulțimea variabilelor exogene este mixtă modelul este de analiză de covarianță.

Analiza regresiei. În modelul de analiză a regresiei relația dintre Y şi $\boldsymbol{X}$ este presupusă liniară, $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ unde: $\boldsymbol{y} = (y_1, y_2, ..., y_n)'$, reprezintă vectorul observațiilor asupra variabilei dependente Y, $\boldsymbol{X} = \{x_{ij}, x_{i0} = 1\}_{i=1}^{n}{}_{j=0}^{p}$, este matricea observațiilor asupra variabilelor explicative, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ reprezintă vectorul coeficienților iar $\boldsymbol{\varepsilon} = (\varepsilon_1,..., \varepsilon_n)'$, este vectorul erorilor/reziduurilor. Pentru evaluarea coeficienților necunoscuți ai modelului, inclusiv a reziduurilor ε_i, se dispune de un sistem de n ecuații liniare având $n+p+1$ necunoscute. Sistemul admite o infinitate de soluții; o soluție posibilă $\boldsymbol{b} = (b_0, b_1, ..., b_p)$ va trebui să minimizeze global mulțimea distanțelor la modelul liniar conform cu un anumit criteriu de definit; sunt aleşi acei vectori $\boldsymbol{b}$ care minimizează mulțimea valorilor $\{e_i\}_{i=1}^{n}$, unde $e_i = y_i - (b_0 + b_1x_{i1} + ... + b_px_{ip})$.

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 funcției de regresie liniară multiplă presupune determinarea coeficienților $b_0, b_1, ..., b_p$ prin metoda celor mai mici pătrate pornind de la observațiile $\{ y_i, x_{i0} = 1, x_{i1}, ..., x_{ip} \}_{i=1}^{n}$. Se presupune că variabilele sunt centrate, ceea ce implică $b_0 = 0$; coeficienții funcției de regresie liniară multiplă sunt $\boldsymbol{b} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$. Căutarea lui $\boldsymbol{y}$ sub forma unei combinații liniare de $\boldsymbol{x_i}$ se reduce la a defini $\tilde{\boldsymbol{y}}$ într-un subspațiu V_X generat de variabilele explicative.

Metoda ajustării celor mai mici pătrate se reduce la aproximarea lui $\boldsymbol{y}$ prin proiecția sa ortogonală $\tilde{\boldsymbol{y}}$, pe V_X înlocuindu-l pe $\boldsymbol{b}$. Se va obține $\tilde{\boldsymbol{y}} = \boldsymbol{Xb} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} = P_X\boldsymbol{y}$, unde P_X este operatorul proiecției ortogonale pe V_X. Lungimile în R^n pot fi interpretate în termeni de dispersie deoarece $(1/n)\sum_{i=1}^{n} y_i^2 = (1/n)\sum_{i=1}^{n}(y_i - \tilde{y})^2 + (1/n)\sum_{i=1}^{n}\tilde{y}_i^2$, unde $(1/n)\sum_{i=1}^{n} y_i^2$, este dispersia totală, $(1/n)\sum_{i=1}^{n}(y_i - \tilde{y})^2$, este dispersia reziduală şi $(1/n)\sum_{i=1}^{n}\tilde{y}_i^2$ reprezintă dispersia explicată.

Pentru o evaluare globală a calității aproximării se definesc: coeficientul de corelație multiplă, $R = cor(\boldsymbol{y}, \tilde{\boldsymbol{y}}) = cor(\boldsymbol{y}, \boldsymbol{Xb})$ şi coeficientul de determinare $R^2 = \Sigma_{i=1}^{n}\tilde{y}_i^2 / \Sigma_{i=1}^{n} y_i^2$ (adică dispersia explicată împărțită la dispersia totală) sau $R^2 = \boldsymbol{y'X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'y} / \boldsymbol{y'y}$ (în funcție de datele inițiale). Dacă $R^2 = 1$ atunci $\tilde{y}_i = y_i \ (\forall)i = 1 \div n$ adică modelul liniar ajustează perfect datele.

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

Din punctul de vedere al prospectorului de date aspectele cele mai interesante privesc semnificațiile statistice ale coeficienților de regresie, adecvarea modelului regresiei multiple la datele observate, studiul reziduurilor, observațiilor aberante şi influenței observațiilor asupra rezultatelor, stabilizarea coeficienților de regresie şi tehnicile de obținere de coeficienți stabili precum şi metodele de selecție a variabilelor ($\boldsymbol{y}$ se „explică” doar prin $q << p$ predictori) pentru a micşora numărul de predictori, a creşte viteza de calcul şi a obține 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 observații asupra variabilei continue Y observată în conjuncție cu cele p variabile nominale $\{X^k\}_{k=1}^{p}$ având, respectiv, modalitățile $\{\tau_k\}_{k=1}^{p}$.

Matricea variabilelor explicative, $\boldsymbol{X}$, se prezintă sub forma $[\boldsymbol{X_1} \cdot \cdots \cdot \boldsymbol{X_k} \cdot \cdots \cdot \boldsymbol{X_p}]$ adică un tablou disjunctiv complet. Pentru fiecare submatrice $\boldsymbol{X_k}$ suma coloanelor este egală cu vectorul $\boldsymbol{1_n}$ existând p relații liniare între coloanele lui $\boldsymbol{X}$. Sistemul de ecuații normale are o infinitate de soluții, toate soluțiile duc la acelaşi vector $\tilde{\boldsymbol{y}}$ care este proiecția lui $\boldsymbol{y}$ pe V_X, dar coeficienții $\boldsymbol{b}$ nu sunt unici. Pentru a obține o estimație unică b, trebuie impuse p restricții liniare privind codificările variabilelor calitative. Cea mai des utilizată restricție este ca suma coeficienților 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 diferența dintre ele şi coloana suprimată. Matricea $\tilde{\boldsymbol{X}}$, a variabilelor explicative astfel recodate este de rang maxim, $rang(\tilde{\boldsymbol{X}} = \sum_{k=1}^{p}(m_k - 1)$.

Pentru exemplificare, în cazul în care se dispune de două variabile nominale A şi B, numite factori, având I, respectiv, J modalități, numite nivele, analiza dispersională cu doi factori cu interacțiune se reduce la a efectua regresia variabilei $\boldsymbol{y}$ cu matricea de condiție $\tilde{\boldsymbol{X}} = [\boldsymbol{1} \cdot \tilde{\boldsymbol{X}}_1 \cdot \tilde{\boldsymbol{X}}_2 \cdot \tilde{\boldsymbol{X}}_{12}]$ cu $rang(\tilde{\boldsymbol{X}}_1)=J$; $rang(\tilde{\boldsymbol{X}}_2)=K$; $rang(\tilde{\boldsymbol{X}}_{12})=JK$, unde $\tilde{\boldsymbol{X}}_1$ şi $\tilde{\boldsymbol{X}}_2$ sunt matricile indicator reduse ale celor doi factori A şi B iar $\tilde{\boldsymbol{X}}_{12}$ este matricea interacțiunilor corespunzând celor JK combinații ale

nivelelor lui A şi B. In această situaţie modelul liniar devine $\boldsymbol{y} = \boldsymbol{\mu 1} + \tilde{\boldsymbol{X}}_1\boldsymbol{\alpha} + \tilde{\boldsymbol{X}}_2\boldsymbol{\beta} + \tilde{\boldsymbol{X}}_{12}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ şi 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.

O anumită prudenţă se impune, totuşi, 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 $M(y_i)$ (numită funcţie de legătură), $a_i = \ell(M(y_i))$ în comparaţie cu modelele liniare obişnuite în care $a_i = M(y_i)$; 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.

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

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 şi 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, $\check{R}$, 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, $\check{S}$, numite suprafeţe de decizie. Dacă suprafeţele de decizie sunt hiperplane H, clasele se zic liniar separabile. Suprafeţele de decizie pot fi descrise cu ajutorul unei mulţimi $G = \{g\}$ de funcţii numite funcţii de discriminare sau funcţii de decizie. Funcţia de discriminare g ataşează fiecare individ $\boldsymbol{x}$ unei regiuni $\check{R}$, 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 asociaţi uneia sau alteia dintre clasele stabilite.

Pentru rezolvarea problemelor de discriminare sunt stabilite reguli de decizie (sau de afectare) şi 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 $g_k(\boldsymbol{x}) : R^p \to R$, $k \in [1, q]$, ecuaţia suprafeţei de decizie ce separă clasa A_k de toate celelalte clase este: $g_k(\boldsymbol{x}) = 0$.

Pentru fiecare clasă A_k, $[\boldsymbol{x} \in A_k] \to [g_k(\boldsymbol{x}) > 0]$. 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 $\check{R}_k$, corespunzătoare clasei A_k, este: $\check{R}_k = \{\boldsymbol{x} \in R^p \mid [g_k(\boldsymbol{x}) > 0] \wedge [g_\ell(\boldsymbol{x}) < 0], (\forall)\ell \in [1, q], \ell \neq k\}$.

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 $g_{k\ell}(\boldsymbol{x}) : R^p \to R$ unde $g_{k\ell}(\boldsymbol{x}) = -g_{\ell k}(\boldsymbol{x})$, $(\forall)\boldsymbol{x} \in R^p$. Suprafaţa de decizie corespunzătoare claselor A_k şi A_ℓ are ecuaţia

$g_{k\ell}(x) = 0$, punctele clasei A_k se află de partea pozitivă a suprafeței. Regula de decizie/afectare este: $[\boldsymbol{x} \in A_k] \leftrightarrow [g_{k\ell}(\boldsymbol{x}) > 0]$ $(\forall)\ell \in [1, q]$, $\ell \neq k$. Regiunea de decizie $\check{R}_k$ corespunzătoare clasei A_k este $\check{R}_k = \{\boldsymbol{x} \in R^p \mid g_{k\ell}(x) > 0, (\forall)\ell \in [1, q], \ell \neq k\}$.

3. Există q funcții 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 $\check{R}_k$ este: $\check{R}_k = \{\boldsymbol{x} \in R^p \mid g_k(\boldsymbol{x}) > g_\ell(\boldsymbol{x}), (\forall)\ell \neq k\}$, $k \in [1, q]$. Suprafața de decizie dintre clasele A_k și A_ℓ este dată de ecuația: $g_k(\boldsymbol{x}) = g_\ell(\boldsymbol{x})$, $(\forall)\boldsymbol{x} \in R^p$, $(\forall)k, \ell \in [1, q]$, $\ell \neq k$. Obiectele clasei A_k se află de partea pozitivă a suprafeței de separare.

Pentru prospectorul de date de o mare importanță practică este cazul claselor liniar separabile. Funcțiile afine de decizie pot fi transformate în funcții liniare de decizie.

Dacă g_k este funcția liniară de decizie corespunzând clasei A_k atunci, în conformitate cu cazul 3 de separabilitate, un obiect $\boldsymbol{x}$ este atașat clasei A_k dacă $g_k(\boldsymbol{x}) > g_\ell(\boldsymbol{x})$ $(\forall)\ell \in [1, q]$, $\ell \neq k$. În cazul 3 de separabilitate regiunile de decizie pot fi mărginite de hiperplane sau de porțiuni de hiperplane. Clasarea, prin minimizarea unei funcții criteriu, conduce la o clasă de funcții discriminante liniare. Funcția criteriu luată în considerație este distanța d de la vectorii caracteristică la prototipurile claselor. Un vector $\boldsymbol{x}$ este atașat acelei clase A_k de al cărei prototip $\boldsymbol{g}_k$ vectorul $\boldsymbol{x}$ este mai aproape, adică: $[d(\boldsymbol{x}, \boldsymbol{g}_k) = min_\ell\ d(\boldsymbol{x}, \boldsymbol{g}_\ell)] \rightarrow [\boldsymbol{x} \in A_k]$.

O clasificare echivalentă se obține considerând funcția de decizie $g_k : R^p \rightarrow R$ dată de formula $g_k(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{g}_k - (1/2)\boldsymbol{g}'_k\boldsymbol{g}_k$. Funcția g_k este o funcție afină de decizie și regula de decizie devine $[g_k(\boldsymbol{x}) = max_\ell\ g_\ell(\boldsymbol{x})] \rightarrow [\boldsymbol{x} \in A_k]$. Hiperplanul de separare este ortogonal pe dreapta ce unește prototipurile claselor, pe care o intersectează într-un punct situat la jumătatea distanței dintre prototipuri. Funcția discriminantă cu distanță 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 de clasare sunt dedicate aspectului inferențial al analizei discriminante.

Fie (Ω, κ, p) un câmp de probabilitate. Probabilitatea condiționată a evenimentului $A \in \kappa$ relativ la evenimentul $B \in \kappa$ cu $p(B) > 0$, este $p_B : \kappa \rightarrow R$ cu $p_B(A) \equiv p(A|B) = p(A \cap B) / p(B)$.

Dacă $\{A_i\}_{i \in I} \subset \kappa$ formează un sistem complet de evenimente atunci are loc următoarea egalitate (*formula lui Bayes a probabilității 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 probabilitățile à priori și $\{p(B|A_i)\}$ probabilitățile à posteriori.

Funcția de repartiție a variabilei aleatoare X condiționată de evenimentul $A \in K$ cu $p(A) > 0$ este funcția $F_A : R \rightarrow [0, 1]$, $F_A(x) \equiv F(x|A) = p(X<x|A)$.

Densitatea de repartiție a variabilei aleatoare X condiționată de evenimentul $A \in K$ cu $p(A) > 0$ este funcția $f(\bullet|A) : R \rightarrow R$ pentru care $F(x|A) = \int_{-\infty}^{x} f(t \mid A)\ dt$.

De asemenea: $f(x|A) = F'(x|A)$ aproape peste tot; $p(A|X=x) = p(A)f(x|A) / f(x)$.

În termenii teoriei statistice a deciziei problema de discriminare se formulează astfel:

Dându-se,

- m grupe (sau populații), $\{\Pi_k\}_{k=1}^m$, specificate prin distribuțiile lor de probabilitate

  $p_k(\boldsymbol{x}) = p(X=\boldsymbol{x}|\boldsymbol{x} \in \Pi_k)$ cu $k = 1 \div m$;

- m probabilități à priori $\{q_k\}_{k=1}^m$, ca un individ să provină din populațiile Π_k, formând un sistem complet de probabilități ($\sum_{k=1}^m q_k = 1$);

- $\mathcal{E} \subset R^p$, spațiul observațiilor asupra a p variabile aleatoare

$\boldsymbol{X} = \{X^j\}^p_{j=1}$ (predictori);

- $\{C(j|k)\}^m_{k,j=1}$, costurile erorilor de clasare (costul clasării unui individ, provenind din populația Π_k în populația Π_j , $j \neq k$);

să se găsească o partiție $\mathcal{P}$ a spațiului $\mathcal{E}$ astfel încât $\sum^m_{k=1} q_k \sum^m_{j=1,\, j\neq k} C(j|k) p(j|k,\ \mathcal{P})$ să fie minimă, unde:

$\mathcal{P} = \{ \check{R}_k\}^m_{k=1}$, $\cup^m_{k=1} \check{R}_k = \mathcal{E}$, $\check{R}_k \cap \check{R}_j = \varnothing$ $(\forall)k, j = 1 \div m, j \neq k$

$\{ p(j|k,\ \mathcal{P}) = \int_{\check{R}j} p_k(\boldsymbol{x}) d\boldsymbol{x}\}^m_{j=1}{}^m_{k=1,k\neq j}$ } sunt probabilitățile de eroare pentru partiția $\mathcal{P}$.

Regula Bayes pentru distribuții cunoscute. Se presupun cunoscute probabilitățile à priori $\{q_k\}^m_{k=1}$ şi distribuțiile de probabilitate $\{p_k\}^m_{k=1}$.

Fie $Y = \{k\}^m_{k=1}$ mulțimea etichetelor claselor şi fie $p_Y(\ell) = \Sigma^m_{k=1} q_k \delta_k(\ell)$ distribuția de probabilitate pe Y, unde $\delta_k(\ell)$ este funcția Dirac ($\delta_k(\ell) = 1$ dacă $\ell = k$ şi $\delta_k(\ell) = 0$ în rest):

- se numeşte plasator o funcție $c : \mathcal{E} \to Y$ ce estimează clasa lui $\boldsymbol{x}$, după ce $\boldsymbol{x} \in \mathcal{E}$ a fost observat $c(\boldsymbol{x}) = \ell$;
- probabilitatea de misclasare pentru clasa k este: $pmc(k) = p[\{c(\boldsymbol{x}) \neq k \,|\, \{\boldsymbol{x} \in \Pi_k\}\}]$;
- funcția de pierdere discretă pentru plasatorul c față de clasa k este notată cu $fpd(c(\boldsymbol{x}), k)$;
- riscul funcțional al plasatorului c este

$$rf(c) \equiv M_\mu[fpd(c(\boldsymbol{x}), k)] = \sum^m_{j=1} q_j pmc(j) = \sum^m_{j=1} q_j \sum^m_{k=1,k\neq j} \int_{\check{R}j} p_k(\boldsymbol{x}) dx$$

  deoarece, distribuția de probabilitate pe $\mathcal{E} \times Y$ este, din construcție, $\mu(\boldsymbol{x},\ k) = q_k\, p_{\ell(x)}(\boldsymbol{x})$ unde cu $\ell(\boldsymbol{x}) \in Y$ s-a notat clasa lui $\boldsymbol{x}$;
- dacă se consideră costurile misclasării $\{C(j|k)\}^m_{k,j=1}$ egale cu 1 atunci un plasator va fi optim dacă minimizează $rf(c) = \sum^m_{k=1} q_k \sum^m_{j=1j\neq k} C(j|k) p(j|k,\ \mathcal{P})$ adică exact funcționala din enunțul problemei de clasare;
- dacă $\boldsymbol{X} = \boldsymbol{x}$ probabilitatea à posteriori a clasei k este $p(k|\boldsymbol{x}) = q_k\, p_k(\boldsymbol{x}) / \sum^n_{k=1} q_k\, p_k(\boldsymbol{x}))$.

Partiția lui $\mathcal{E}$ care minimizează riscul funcțional $rf(c)$ este $\mathcal{P} = \{ \check{R}_k\}^m_{k=1}$, unde regiunile de decizie $\check{R}_k = \{ \boldsymbol{x} \in \mathcal{E} \,|\, \sum^m_{j=1j\neq k} q_j p_j(\boldsymbol{x}) \leq \sum^m_{j=1j\neq \ell} q_j p_j(\boldsymbol{x}),\ (\forall)\ell \in [1, m],\ \ell \neq k \}$ sunt numite regiuni de decizie Bayes şi 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 funcțional este notat cu $c_B(\boldsymbol{x})$. Plasatorul $c_B(\boldsymbol{x})$ se numeşte plasator Bayes, riscul funcțional pe care acesta îl minimizează se numeşte risc (sau eroare) Bayes, iar partiția $\mathcal{P}$, 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* $p((p_j(\boldsymbol{x})/p_\ell(\boldsymbol{x}))=b|\boldsymbol{x}\in\Pi_k)=0$, $(\forall)j, k, \ell = 1 \div m$, $\ell \neq j$ *şi* $0 \leq b < \infty$, *atunci clasa procedurilor bayesiene este minimală şi completă*.

Regula Bayes pentru distribuții cunoscute permite să se construiască o procedură de clasare cu proprietăți de optimalitate dar aplicabilitatea practică directă este redusă deoarece, în realitate, cel puțin distribuțiile $\{ p_k\}^m_{k=1}$ 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 densitățile 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 $[X\in\Pi_k]\to[X \sim N(\boldsymbol{\mu}_k, \boldsymbol{V})]$, unde $\boldsymbol{\mu}_k \in M_{p\times 1}(R)$ este vectorul medie şi $\boldsymbol{V} \in M_{p\times p}(R)$ este matricea de varianță-covarianță:

- regiunea de clasificare în Π_1, şi anume $\check{R}_1$, este mulțimea punctelor $\boldsymbol{x} \in R^p$ pentru care raportul densităților $p_1(\boldsymbol{x}) / p_2(\boldsymbol{x}) \geq c$, cu c o constantă convenabil aleasă. Condiția de definire a lui $\check{R}_1$ revine la: $f(\boldsymbol{x}) \equiv \boldsymbol{x}'\boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) + (-1/2)(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)' \boldsymbol{V}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \geq ln\, c$;

- dacă $\{\Pi_k\}^2_{k=1}$ sunt populații multidimensionale, normal distribuite, de medie $\boldsymbol{\mu_i}$ și cu matricea $\boldsymbol{V}$ de varianță-covarianță, comună, atunci cele mai bune regiuni de clasificare sunt date de: $\check{R}_1 \equiv f(\boldsymbol{x}) \geq \ln c$ și $\check{R}_2 \equiv f(\boldsymbol{x}) < \ln c$;
- dacă probabilitățile à priorice q_1 și 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 suprafața de separare a celor două regiuni este hiperplanul H: $(\boldsymbol{g_1} - \boldsymbol{g_2})'(\boldsymbol{x} - (1/2)(\boldsymbol{g_1} + \boldsymbol{g_2})) = 0$, unde $\boldsymbol{g_k} = \boldsymbol{V^{-1}\mu_k}$ este prototipul populației Π_k iar clasificatorul obținut este un clasificator cu distanță minimă;
- dacă probabilitățile à priorice nu sunt cunoscute atunci constanta $C = \ln c$ va fi aleasă 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 distanța Mahalanobis dintre cele două populații iar $\Phi(x) = \int_{-\infty}^{x} (1/\sqrt{2\pi})e^{\varphi(t)}dt$ cu $\varphi(t) = -(t^2/2)$, este funcția de repartiție a variabilei aleatoare Gauss-Laplace.

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

Dacă $\boldsymbol{x_1}^{(i)}, ..., \boldsymbol{x_{ni}}^{(i)} \in N(\boldsymbol{\mu_i}, \boldsymbol{V})$, $i \in \{1, 2\}$ sunt două selecții bernoulliene atunci estimatorii $\bar{\boldsymbol{x}}_i = (1/n_i)\sum_{j=1}^{ni}\boldsymbol{x}^{(i)}_j$ și $((n_1-1) + (n_2-1))\boldsymbol{S} = (n_1 + n_2 - 2)\boldsymbol{S} = \sum_{i=1}^{2}\sum_{j=1}^{ni}(\boldsymbol{x}^{(i)}_j - \bar{\boldsymbol{x}}_i)(\boldsymbol{x}^{(i)}_j - \bar{\boldsymbol{x}}_i)'$ sunt estimatori nedeplasați, de verosimilitate maximă, ai lui $\boldsymbol{\mu_i}$, și $\boldsymbol{V}$. Pentru selecții suficient de mari folosirea estimațiilor în locul valorilor exacte implică erori mici.

Substituind parametrii estimați în relațiile de definiție ale regiunilor de decizie se obține: $\check{R}_1 \equiv f(\boldsymbol{x}) \geq \ln c$ și $\check{R}_2 \equiv f(\boldsymbol{x}) < \ln c$, unde $f(\boldsymbol{x}) = \boldsymbol{x}'\boldsymbol{S^{-1}}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)}) - (1/2)(\bar{\boldsymbol{x}}^{(1)} + \bar{\boldsymbol{x}}^{(2)})'\boldsymbol{S^{-1}}(\bar{\boldsymbol{x}}^{(1)} - \bar{\boldsymbol{x}}^{(2)})$.

Dacă se dorește clasificarea selecțiilor reunite ca un tot se utilizează următorii estimatori:

$n = n_1+n_2$, $\bar{\boldsymbol{x}} = (1/n)\sum_{j=1}^{n}\boldsymbol{x_j}$, $[\boldsymbol{x_j} \in \Pi_1] \vee [\boldsymbol{x_j} \in \Pi_2]$; $(n_1 + n_2 + n - 3)\bar{\boldsymbol{S}} = \boldsymbol{S} + \sum_{j=1}^{n}(\boldsymbol{x_j} - \bar{\boldsymbol{x}})(\boldsymbol{x_j} - \bar{\boldsymbol{x}})'$.

$\check{R}_1 \equiv (\bar{\boldsymbol{x}} - (1/2)(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2))'\boldsymbol{S^{-1}}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \geq c$.

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 selecție aleatoare dintr-o populație având densitatea de probabilitate $f(\boldsymbol{x}; \theta)$ cu $\boldsymbol{x} \in \boldsymbol{X}$ și $\theta \in \Theta$. O procedură de inferență frecventistă depinde de funcția de verosimilitate $L(\theta) = \prod_{i=1}^{n} f(x_i; \theta)$, unde θ este necunoscut dar fixat.

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

- dacă $f(\boldsymbol{x}|\theta)h(\theta)$, distribuția comună a lui $\boldsymbol{x}$ și θ, și $m(\boldsymbol{x}) = \int_{\Theta} f(\boldsymbol{x}|\theta)h(\theta)d\theta$, distribuția marginală a lui $\boldsymbol{x}$, sunt cunoscute, atunci distribuția lui θ condiționată de evenimentul $\boldsymbol{X}=\boldsymbol{x}$ sau distribuția à posteriori a lui θ, este: $h(\theta|\boldsymbol{x})=h(\theta|X=\boldsymbol{x}) = f(\boldsymbol{x}|\theta)h(\theta) / m(\boldsymbol{x})$, $m(\boldsymbol{x})>0$, $\boldsymbol{x} \in \mathcal{E}$, $\theta \in \Theta$. $\tilde{\Theta}$;
- dacă $\theta \sim N(\boldsymbol{m}, \boldsymbol{S})$ și $\boldsymbol{x} \sim N(\theta, \boldsymbol{V})$, atunci $h(\theta|\boldsymbol{x})$ este densitatea de probabilitate a unei $N(\mu, \boldsymbol{C})$ cu $\mu = \boldsymbol{S}(\boldsymbol{S} + \boldsymbol{V})^{-1}\boldsymbol{x} + \boldsymbol{V}(\boldsymbol{S} + \boldsymbol{V})^{-1}\boldsymbol{m}$ și $\boldsymbol{C} = \boldsymbol{V}(\boldsymbol{S} + \boldsymbol{V})^{-1}\boldsymbol{S}$;
- dacă $\theta \sim N(\tau, \sigma^2_0)$ și $x \sim N(\theta, \sigma^2_1)$, atunci densitatea à posteriori a lui θ este: $N(\mu, \sigma^2)$, unde $\mu = (x/\sigma^2_1 + \tau/\sigma^2_0)\,(1/\sigma^2_0 + 1/\sigma^2_1)^{-1}$ și $\sigma^2 = (\sigma^2_0\,\sigma^2_1)/(\sigma^2_0 + \sigma^2_1) = (1/\sigma^2_0 + 1/\sigma^2_1)^{-1}$.

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

Fie $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ o selecție bernoullianăasupra unei variabile aleatoare ce depinde de θ şi fie $\tilde{\theta} \equiv \tilde{\theta}(T)$ un estimator a lui θ:

- funcția de pierdere, estimând θ prin $\tilde{\theta}$, este: $L^b(\theta, \tilde{\theta}) \equiv L^b(\theta, \tilde{\theta}(T)) = (\tilde{\theta}(T) - \theta)^2$;
- riscul funcțional este: $R^b(\theta, \tilde{\theta}) = M[L^b(\theta, \tilde{\theta})] = \int_\Delta L^b(\theta, \tilde{\theta}(t)) f(t|\theta) dt$;
- se numeşte risc bayesian: $r^b(\theta, \tilde{\theta}) = \int_\Theta R^b(\theta, \tilde{\theta}) h(\theta) d\theta$;
- se numeşte estimator bayesian: $r^b(\theta, \tilde{\theta}^b) = inf_{\tilde{\theta} \in B}\, r^b(\theta, \tilde{\theta})$, $\tilde{\theta}^b \in B$, unde B este clasa estimatorilor pentru care riscul bayesian este finit;
- în cazul funcției de pierdere „suma pătratelor erorilor" estimatorul bayesian este dat de relația: $\tilde{\theta}^b(t) = \int_\Theta \theta h(\theta|t) d\theta \equiv M[\theta|T(\boldsymbol{x}) = t]$, adică media distribuției 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 şi fie statistica $T = (1/n)\Sigma^n_{i=1} \boldsymbol{x}_i$, care este suficientă pentru θ:

- dacă distribuția a priori a lui θ pe spațiul $\Theta = R$ este $N(\tau, \sigma^2_0)$ cu τ, $\sigma_0 \in R$ dați şi $\sigma_0 > 0$, distribuția à posteriori a lui θ, condiționată de observațiile $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, 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 observațiile efectuate;
- dacă $\sigma_0 > \sigma_1$ rezultă $\mu \approx \bar{\boldsymbol{x}}$, cunoaşterea mediei à priorice τ este de importanță 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 $lim_{n \to \infty} \mu = lim_{n \to \infty} \bar{\boldsymbol{x}}$;
- dacă dispersia inițială este mică, media estimată tinde să rămână în apropierea mediei inițiale τ chiar dacă media empirică $\bar{\boldsymbol{x}}$ diferă considerabil de aceasta;
- dacă raportul a este mic, atunci media şi dispersia à priori au doar o influență redusă asupra estimării parametrilor care sunt determinați aproape exclusiv din datele empirice;
- dacă $T(\boldsymbol{x}) = t$, atunci estimatorul Bayes al mediei unei variabile aleatoare $N(\mu, \sigma^2)$ este: $\tilde{\theta}(t) = \tilde{\theta}_B = (nt/\sigma^2_1 + nt/\sigma^2_0)\,(1/\sigma^2_1 + 1/\sigma^2_0)^{-1}$;
- pentru cazul multidimensional: $\tilde{\theta}_B = \boldsymbol{S}(\boldsymbol{S} + (1/n)\boldsymbol{V})^{-1}\boldsymbol{t} + (1/n)\boldsymbol{V}(\boldsymbol{S} + (1/n)\boldsymbol{V})^{-1}\boldsymbol{m}$.

Fie $\boldsymbol{X} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$ o selecție bernoullianădin populațiile Π_1 şi Π_2:

- dacă $\boldsymbol{X} \in \Pi_1$, atunci densitatea de probabilitate este $f_k(\boldsymbol{x}|\theta)$, $\theta \in \theta_k$ şi densitatea à priorică este $h_k(\theta)$, $k = \{1, 2\}$;
- dacă q_1 şi q_2 sunt probabilitățile à priori ale populațiilor Π_1, şi Π_2, probabilitățile à posteriori sunt: $p(\Pi_k|\boldsymbol{x}) = m_k(\boldsymbol{x})q_k / (m_k(\boldsymbol{x})q_k + m_k(\boldsymbol{x})q_k)$, unde $m_k(\boldsymbol{x}) = \int_{\Theta k} f_k(x|\theta)\, h_k(\theta)\, d\theta$, este densitatea de probabilitate marginală a lui $\boldsymbol{x}$ condiționat de faptul că provine din Π_k;
- procedura bayesiană de discriminare este:

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

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

6. Maşini cu suport vectorial

Maşinile cu suport vectorial reprezintă o clasă de algoritmi de instruire 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 modelului controlându-i complexitatea, mai precis, integrând în estimare numărul de parametri, în acest caz numărul de vectori suport. Ideea de bază al maşinilor cu suport vectorial a fost de a reduce problema discriminării la o problemă, liniară, de căutare a unui hiperplan optimal:

- 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;
- prin căutarea unor suprafețe de separare neliniare;
- prin introducerea unei funcții nucleu în produsul scalar inducând implicit o transformare neliniară a datelor către un spațiu hilbertian, intermediar, de dimensiune mai mare şi în care este rezolvată problema liniară.

Fie Y variabila de explicat, fie $X = \{X^j\}_{j=1}^p$ o variabilă explicativă sau de predicție multidimensională şi fie φ un model pentru Y:

- $X = \{X^j\}_{j=1}^p$ este o variabilă cu valori într-o mulțime $\mathcal{E} \subset R^p$, $\boldsymbol{x} = (x_j)_{j=1}^p \in \mathcal{E}$;
- Y este dicotomică, $\varphi : \mathcal{E} \to B$, $\varphi(\boldsymbol{x}) \in B = \{-1, 1\}$;
- $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ este un eşantion de mărime n şi de lege F necunoscută;
- obiectivul este de a construi o estimare $\tilde{\varphi} : \mathcal{E} \to \{-1, 1\}$ a lui φ astfel încât probabilitatea $\rho(\varphi(X) \neq Y)$ să fie minimă.

Problema revine la a căuta o frontieră de decizie în spațiul $\mathcal{E}$ pentru valorile lui X şi la a găsi un compromis între complexitatea acestei frontiere, respectiv, capacitatea de ajustare a modelului, şi 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: $\rho(\varphi(X) \neq Y) = \rho(Yf(X) \leq 0)$. 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 $\mathcal{E}$, prin funcția $\Phi : \mathcal{E} \to \tilde{H}$ cu valori într-un spațiu $\tilde{H}$, intermediar, înzestrat cu un produs scalar. Această transformare, fundamentală pentru abordarea *SVM*, ia în considerare eventuala neliniaritate a problemei de rezolvat şi 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 $\tilde{H}$, un hiperplan H este definit prin ecuația $\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b = 0$, unde $\boldsymbol{w}$ este un vector ortogonal pe hiperplan, $\boldsymbol{w} \perp H$, 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 $H \equiv (\boldsymbol{w}, b)$ 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: $d(\boldsymbol{x}) = |\langle \boldsymbol{w}, \boldsymbol{x} \rangle + b| / \|\boldsymbol{w}\| = |f(\boldsymbol{x})| / \|\boldsymbol{w}\|$;
- marja hiperplanului are valoarea $2 / \|\boldsymbol{w}\|^2$.

Căutarea hiperplanului separator de marjă maximală revine la rezolvarea problemei (primare) de optimizare cu restricții: $(1/2) min_{\boldsymbol{w}} \|\boldsymbol{w}\|^2 \cdot y_i(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) \geq 1$ $(\forall) i$. Problema duală se obține prin introducerea multiplicatorilor Lagrange. Soluția este furnizată de un punct şa $(\boldsymbol{w}^*, b^*, \lambda^*)$ al lagranjianului $L(\boldsymbol{w}, b, \lambda) = (1/2) \cdot \|\boldsymbol{w}\|^2 - \sum_{i=1}^n \lambda_i (y_i (\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b) - 1)$, punctul şa 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: $W(\boldsymbol{\lambda}) = \sum_{i=1}^{n} \lambda_i - (1/2)\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i\lambda_j y_i y_j \langle \boldsymbol{x_i}, \boldsymbol{x_j}\rangle$. 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, furnizează ecuația hiperplanului optimal:

$$\sum_{i=1}^{n} \lambda^*_i y_i = \langle \boldsymbol{x}, \boldsymbol{x_i}\rangle + b^* = 0 \text{ cu } b_0 = -(1/2)(\langle \boldsymbol{w}^*, \boldsymbol{SV_{clasa+1}}\rangle + \langle \boldsymbol{w}^*, \boldsymbol{SV_{clasa-1}}\rangle).$$

Pentru o nouă observație $\boldsymbol{x}$ prezentată modelului, este suficient să se vadă semnul expresiei $f(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda^*_i y_i \langle \boldsymbol{x}, \boldsymbol{x_i}\rangle + b^*$ 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 π pentru încălcarea restricțiilor, problema de minimizare se reformulează în felul următor:

$$min(1/2) \cdot \|\boldsymbol{w}\|^2 + \pi \sum_{i=1}^{n} \xi_i \cdot y_i\langle \boldsymbol{w}, \boldsymbol{x_i}\rangle + b \geq 1 - \xi_i, (\forall) i \in [1, n].$$

Problema se formulează în aceeași formă duală ca și în cazul separabilității cu o singură diferență: coeficienții λ_i sunt mărginiți de constanta π de control a penalizării. Din punctul de vedere al prospectorului de date parametrul π, 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 $\mathcal{E}$ sunt transformate prin aplicația neliniară $\Phi : \mathcal{E} \rightarrow \tilde{H}$, spațiul $\tilde{H}$ fiind de dimensiune mai mare și înzestrat cu un produs scalar. Formularea problemei de minimizare și soluția sa $f(\boldsymbol{x}) = \sum_{i=1}^{n} \lambda^*_i y_i\langle \boldsymbol{x}, \boldsymbol{x_i}\rangle + b^*$ 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 $\tilde{H}$ cu ajutorul unei funcții $\mathsf{k} : \mathcal{E} \times \mathcal{E} \rightarrow R$, simetrică, numită nucleu (*kernel*), astfel încât: $\mathsf{k}(\boldsymbol{x}, \boldsymbol{x'}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{x'})\rangle$. 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 liniare $\mathsf{k}(\boldsymbol{x'}, \boldsymbol{x''}) = \langle \boldsymbol{x'}, \boldsymbol{x''}\rangle$, polinomiale $\mathsf{k}(\boldsymbol{x'}, \boldsymbol{x''}) = (c + \langle \boldsymbol{x'}, \boldsymbol{x''}\rangle)^d$ sau gaussiene $\mathsf{k}(\boldsymbol{x'}, \boldsymbol{x''}) = e^{-\alpha(x', x'')}$, 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.

7. 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, se caracterizează prin: stări interne, $s \in S$, semnale de intrare $\{x_i\}_{i=1}^{p}$, funcția de tranziție a stărilor $s = h(x_1, \ldots, x_p) = f(\beta_0 + \sum_{j=1}^{p} \beta_j x_j)$. Valorile coeficienților $\{\beta_j\}_{j=0}^{p}$ sunt estimate într-o fază de instruire 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 $\{x_j\}_{j=1}^{p}$ 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, ieşirea 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 instruire.

Pentru un eşantion de instruire $\{(x^1_i, \ldots, x^p_i ; y_i)\}^n_{i=1}$ construit din n observații asupra a p variabile explicative $\{X^j\}^p_{j=1}$ şi a unei variabile de explicat Y, instruirea constă în estimarea vectorului de parametri $\boldsymbol{\beta}$ rezolvând o problemă a celor mai mici pătrate:

$\tilde{\boldsymbol{\beta}} = min_{\boldsymbol{b}}\, Q(\boldsymbol{b})$, $\blacksquare$ $Q(\boldsymbol{b}) = (1/n)\sum^n_{i=1}(y_i - \varphi(x^1_i, \ldots, x^p_i; (\boldsymbol{b}))\,)^2$.

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 creştere a erorii şi deci pentru a-l descreşte pe Q este suficientă o deplasare în sens contrar. Pornind de la erorile observate pe ieşiri, formula retropropagării erorii furnizează expresia erorii atribuite fiecărei intrări, de la stratul de ieşire 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 eşantionului de instruire tinde către infinit.

În practică, prospectorul de date se confruntă cu o serie de opțiuni privind, în principal, controlul suprainstruirii: alegerea unor parametri (limitarea numărului de neuroni, limitarea duratei de instruire, creşterea coeficientului de penalizare a normei parametrilor); alegerea modului de estimare a erorii (pe eşantionul de test sau validare încrucişată).

8. Metoda segmentării

Metoda segmentării este o metodă complementară de rezolvare a problemelor de discriminare şi de regresie prin împărțirea progresivă a eşantionului de observații într-un arbore de decizie binară.

Fie $\boldsymbol{y}$ variabila privilegiată, discretă, cu q modalități, $\{k\}^q_{k=1}$, care este explicată prin variabilele, cantitative sau calitative, $\{X^j\}^p_{j=1}$, şi fie $\{\boldsymbol{x_i}\ ; y_i)\}^n_{i=1} \equiv \{\{x^j_i\}^p_{j=1}; y_i)\}^n_{i=1}$ eşantionul 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 $\boldsymbol{y}$ şi defineşte o împărțire a eşantionului în două submulțimi 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 împărțirea succesivă a eşantionului în câte două submulțimi rezultă un arbore de decizie binară în care se disting: segmente intermediare, segmente terminale, ramuri ale unui segment, arborele binar complet, A_{max}, şi 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 față de variabilă.

Fazele de construire ale arborelui de decizie binară sunt: stabilirea, pentru fiecare nod, a mulțimii diviziunilor admisibile; definirea unui criteriu de selecționare 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.

Inițial, există un singur segment conținând toți indivizii $\boldsymbol{x_i}$, $i = 1 \div n$. Sunt examinate, secvențial, 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 selecționată 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 *s*e aplică iterativ fiecărui segment descendent obținut şi 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 condiționată de apartenență la grupul G_r, $r \in \{1, 2, ..., q\}$ a mulțimii

observațiilor din nodul a. Impuritatea unui nod, a, este o funcție nenegativă de $\{p(r|a)\}_{r=1}^{q}$, care verifică următoarele condiții: este maximală când probabilitățile de apartenență la diferite grupuri sunt egale între ele: $p(r|a) = 1$, $(\forall)r$; este nulă dacă nodul conține observații aparținând unui singur grup: $p(r|a) = 1$ şi $p(s|a) = 0$, $(\forall)r,s\ s \neq r$; este o funcție simetrică de probabilitățile $p(r|a)$. Funcțiile de impuritate, cele mai des utilizate, sunt: $\imath(a) = -\sum_{r=1}^{q} p(r|a)\ ln(p(r|a))$, funcție derivată din noțiunea de entropie Shannon şi indicele de diversitate Gini $j(a) = -\sum_{r\neq s} p(r|a)\ p(s|a)$.

Fie ∂ o diviziune admisibilă care împarte/divide nodul a în segmentele t_s şi t_d cu probabilitățile: $p_s \equiv p(t_s|a) = p(t_s) / p(a)$ şi respectiv $p_d \equiv p(t_d|a) = p(t_d) / p(a)$. Reducerea impurității 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 impurității. Cea mai „bună" diviziune este $\tilde{\partial}^j = argmax_{m\in\partial j}\ \Delta_i(\partial^j_m, t)$ adică aceea pentru care reducerea impurității este maximă, unde ∂^j este mulțimea diviziunilor admisibile ale variabilei X^j. Pe mulțimea $\{X^j\}_{j=1}^{p}$, a 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 preț de misclasare $\gamma(s/r)$, $s, r = 1, 2, ..., q$ costul misclasării fiind $\sum_{r=1}^{q} \gamma(s/r)\ p(r|a)$.

Un nod a va fi asignat acelei clase $\tilde{s}$ pentru care $\tilde{s} = min_{1\leq s\leq q} \sum_{r=1}^{q} \gamma(s/r) p(r|a)$. Dacă minimul este atins pentru cel puțin 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 mulți reprezentanți în ea.

Costul misclasării unei observații aparținând nodului a este: $c(a) = min_{1\leq s\leq q} \sum_{r=1}^{q} \gamma(s/r) p(r|a)$. 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: $rea(A) = \sum_{a\in\acute{A}} C(a) = \sum_s \sum_{a\in\acute{A}(s)} \sum_r \gamma(s/r)\ p(r|a)\ p_r = \sum_s \sum_r \gamma(s/r)(n_{sr}/n)$, unde $\acute{A}$ este mulțimea nodurilor terminale ale lui A, $\acute{A}(s)$ este mulțimea 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 clasați în clasa s, $s \neq r$.

Un subarbore al lui A_{max} este optimal („cel mai bun") dacă numărul de segmente terminale conținute şi riscul erorii de afectare sunt minime şi, în plus, furnizează o estimație corectă a erorii teoretice de clasare. Pentru selecția subarborelui optimal se împarte eşantionul inițial într-un eşantion de instruire şi un eşantion de testare. Pornind de la eşantionul de instruire se construieşte arborele A_{max}. Operația 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 toți subarborii având acelaşi număr de segmente terminale, $ea(A_h) = min_{A\in S_h} ea(A)$, unde S_h este mulțimea subarborilor lui A_{max} cu h segmente terminale. Se selectează din şirul de arbori optimali subarborele $\tilde{A}$ cu eroarea teoretică minimă dată de relația: $et(\tilde{A}) = min_{1\leq h\leq H} et(A_h)$. Eroarea teoretică se estimează după formula $\tilde{et}(A) = \sum_{t\in A} \tilde{R}_t$, unde avem: $\tilde{R}_t = (\tilde{n}_t / \tilde{n}) \times \tilde{s}^2_t$, $\tilde{n}$ este volumul eşantionului de test, $\tilde{n}_t$ este numărul de indivizi din eşantionul de test aparținând segmentului t, $\tilde{s}^2_t$ este dispersia de selecție a variabilei $\boldsymbol{y}$ în interiorul segmentului t, $\tilde{s}^2_t = (1 / \tilde{n}_t) \sum_{i=1}^{|t|} (y_i - \tilde{\bar{y}})$, $\tilde{\bar{y}}$ este media de selecție în interiorul segmentului t, $|t| = card(t)$.

Deşi cea mai bună diviziune, $\tilde{\partial}$, 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 şi alte diviziuni (echi-reductive, echi-divizante), aproximativ la fel de bune, dar foarte importante la nivelul interpretării.

9. Metode de agregare a modelelor

Agregarea (sau combinarea) unui număr mare de modele permite ameliorarea ajustării modelelor evitându-se, totodată, supraajustarea acestora şi se bazează pe două tipuri de strategii de

agregare: aleatoare (*bagging*) şi 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 $X=\{X^j\}_{j=1}^{p}$ variabilele explicative, fie $\varphi(X)$ un model funcţie de X şi fie $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ un eşantion de lege F. Speranţa, $M_F(\tilde{\varphi}_z)$, 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 $\{\boldsymbol{z}_\kappa\}_{\kappa=1}^{K}$, şi se construieşte familia de modele $\{\varphi_{\boldsymbol{z}\kappa}\}_{\kappa=1}^{K}$.

Estimarea medie va fi:

$$\tilde{\varphi}_{\boldsymbol{K}}(\bullet) = \begin{cases} M_F(\tilde{\varphi}_{z\kappa}) = (1/K)\sum_{\kappa=1}^{K} \tilde{\varphi}_{z\boldsymbol{k}}(\bullet), \text{ dacă variabila de explicat } Y \text{ este cantitativă} \\ argmax_{1\le\tau\le q} \left| \{ \kappa \mid \tilde{\varphi}_{z\boldsymbol{k}}(\bullet) = \tau,\ \kappa = 1 \div K \} \right|, \text{ dacă } Y \text{ este calitativă} \end{cases}$$

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 eşantioane independente, $\boldsymbol{z}_\kappa$, ar necesita, în general, prea multe date şi ele sunt înlocuite prin K eşantioane bootsrap , $\boldsymbol{z}^{*}_{\kappa}$, obţinute, fiecare, prin n extrageri cu înlocuire conform legii empirice $\tilde{F}$. În fiecare iteraţie κ ($\kappa = 1 \div K$), se extrage eşantionul bootstrap, $\boldsymbol{z}^{*}_{\kappa}$ şi se calculează $\tilde{\varphi}_{z^{*}\kappa}(\boldsymbol{x})$ pe acest eşantion. 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 îmbunătăţire a bagging-ului se poate obţine prin adăugarea unei randomizări. Obiectivul este de mări independenţa arborilor de agregare prin intervenţia hazardului în alegerea variabilelor implicate în modele. În fiecare iteraţie κ ($\kappa = 1 \div K$): se extrage un eşantion bootstrap $\boldsymbol{z}^{*\kappa}$ şi se estimează un arbore pe $\boldsymbol{z}^{*\kappa}$ prin randomizarea variabilelor (căutarea fiecărui nod optimal este precedată de selecţia aleatoare a unei submulţimi de $q \le p$ predictori). In final, $\tilde{\varphi}_{\boldsymbol{K}}(\boldsymbol{x}) = (1/K)\sum_{\kappa=1}^{K} \tilde{\varphi}_{z^{*}\kappa}(\boldsymbol{x})$ sau $\tilde{\varphi}_{\boldsymbol{K}}(\boldsymbol{x})$ = rezultatul votului. Faţă de bagging, în cazul „pădurilor aleatoare" (*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 limitaţi la o singură ramificaţie riscă să fie foarte asemănători (puternic corelaţi) implicând, aceleaşi, câteva variabile care apar ca fiind cele mai explicative. În fiecare etapă de construcţie a unui arbore, selectarea aleatoare a unui număr redus de predictori potenţiali creşte semnificativ variabilitatea având în mod necesar alte variabile. Fiecare model de bază este în mod evident mai puţin 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. Informaţiile relevante sunt obţinute prin calcul şi prin reprezentarea grafică a unor indici, proporţionali cu importanţa fiecărei variabile din modelul agregat adică cu participarea acesteia la regresie sau discriminare. Aceste informaţii sunt cu atât mai utile cu cât variabilele sunt mai numeroase. Pentru a evalua importanţa unei variabile prospectorul de date utilizează criterii precum: frecvenţa cu care apare fiecare variabilă în arborii pădurii, *MDA* (*Mean Decrease Accuracy*) sau *MDG* (*Mean Decrease Gini*).

Strategii adaptive. Boosting-ul adoptă acelaşi principiu general ca şi 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 priveşte 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 îşi 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 şi fie δ funcţia de discriminare cu valori în $\{-1, 1\}$. Pentru estimarea primului model ponderile w_i ale fiecărei observaţii sunt iniţializate la $1/n$, în continuare aceste ponderi evoluează la fiecare iteraţie adică pentru fiecare nouă estimare. Importanţa, w_i, a unei observaţii rămâne neschimbată dacă observaţia este bine clasată, dacă observaţia nu este bine clasată w_i creşte proporţional cu deficitul de ajustare al modelului. Agregarea finală a previziunilor, $\sum_{\kappa=1}^{K} c_\kappa \delta_\kappa(\boldsymbol{x})$, este o combinaţie ponderată a calităţilor de ajustare ale fiecărui model. Valoarea absolută a sa, numită marje, este proporţională cu încrederea care poate fi acordată semnului său care furnizează rezultatul previziunii.

Fie $\boldsymbol{z} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ un eşantion şi $\boldsymbol{x}$ individul de previzionat. Se iniţializează $\boldsymbol{w}_1$, vectorul de ponderi: $w_{1,i} = 1/n$. În fiecare iteraţie κ ($\kappa = 1 \div K$): mai întâi se estimează δ_κ pe eşantionul $\boldsymbol{z}_\kappa$ ($\boldsymbol{z}$ ponderat cu $\boldsymbol{w}_\kappa$); apoi se consideră vectorul $\boldsymbol{Q}_\kappa = \{Q\kappa,i\}_{i=1}^{n}$, unde $Q\kappa,i$ este un indice de misclasare ($Q\kappa,i = 1$ dacă $\delta_\kappa(x_i) \neq y_i$ şi $Q\kappa,i = 0$ dacă $\delta_\kappa(x_i) = y_i$) cu ajutorul căruia se estimează eroarea de previziune $\tilde{\varepsilon}_P = (\sum_{i=1}^{n} w_i\, Q\kappa,i) / (\sum_{i=1}^{n} w_i)$; se calculează $c_\kappa = log((1 - \tilde{\varepsilon}_P) / \tilde{\varepsilon}_P)$; se calculează noile ponderi: $w_{\kappa+1,i} = w_{\kappa,i}\, exp[-c_\kappa Q\kappa,i]$, $i = 1 \div n$. În final, rezultatul este: $\tilde{\varphi}_K(\boldsymbol{x}) = sign[\sum_{\kappa=1}^{K} c_\kappa\, \delta_\kappa(\boldsymbol{x})]$.

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.

10. Concluzii

Practica de a obţine din date cunoştinţe valoroase şi utile pentru susţinerea activităţilor decizionale, denumită tot mai frecvent *data science*, este în continuă şi rapidă dezvoltare pentru a face faţă provocărilor de prelucrare a seturilor uriaşe de date (structurate, nestructurate sau semi-structurate generate de dispozitive inteligente, telefoane mobile, web, mass-media sau reţele sociale), *big data*.

Informatica decizională utilizează statistica descriptivă, pentru date cu mare densitate în informaţie, pentru a măsura fenomene, a detecta tendinţe, etc. în timp ce *big data* utilizează statistica inferenţială, pentru date cu slabă densitate în informaţie, ale căror volume, foarte mari, permit inferenţe ale legilor conferindu-le capacităţi predictive (cu limitele acestor inferenţe).

Pentru prospectarea datelor şi interpretarea rezultatelor *data scientist*, specializat de obicei pe un anumit domeniu (marketing, medicină, securitate, fraudă, finanţe, etc.), se bazează pe expertize din statistică, instruire, optimizare, procesare de semnale, regăsire de informaţii sau procesare a limbajului natural.

Având o pregătire de bază în matematică şi statistică, noul *data scientist* poate privi cu seninătate sosirea valului sau tsunami-ului *Big Data*.

Activitatea informatică din amonte (permanent reînnoită de evoluţia rapidă a tehnologiilor) este importantă, pentru a stoca datele şi a face executabile metodele dar, conceptual, matematica necesară modelelor respective a luat deja în considerare mărimi şi dimensiuni infinite în spaţii hilbertiene.

Înzestrat cu acest „instrumentar” durabil, *data scientist* poate deci aborda şi susţine, cu şanse de succes, cercetările emergente.

BIBLIOGRAFIE

1. **BANCIU, D.; COARDOŞ, D.; LEPĂDATU, C-I.; LEPĂDATU, C.:** Enhancement of the Retrospective National Bibliography of the Romanian Book through the Application of the Informational Technologies, Proceedings of BIBLIO 2011 „Innovation en

bibliotheque/Innovation within libraries", Editura Universităţii Transilvania din Braşov, 2011, pp. 131-142.

2. **BESSE, P.; LAURENT, B.:** Apprentissage Statistique: modélisation, prévision et data mining, Institut National des Sciences Appliquées de Toulouse, 2014, 159 p.
3. **CIUREA, C.; DUMITRESCU, G.; LEPĂDATU, C.:** The impact analysis of implementing virtual exhibitions for mobile devices on the access to national cultural heritage, Proceedings of 2nd International Conference Economic Scientific Research – Theoretical, Empirical and Practical Approaches, ESPERA 2014, Bucharest, Romania.
4. **COARDOŞ, D.; COARDOŞ, V.; LEPĂDATU, C-I.; LEPĂDATU, C.:** Support Systems for Libraries Based on Business Intelligence Tools, 2008 IEEE International Conference on Intelligent Computer Communication and Processing - Digital Libraries Workshop, Cluj Napoca, August 2008.
5. **COARDOŞ, D.; COARDOŞ, V.; LEPĂDATU, C-I.; LEPĂDATU, C.:** Integrated On-line System for Management of the National Retrospective Bibliography – SIMBNR, 2009 IEEE International Conference on Intelligent Computer Communication and Processing - Workshop on Digital Libraries, e-Content Management and e-Learning", Cluj Napoca, August 2009.
6. **DUMITRESCU, G.; FILIP, F.-G.; IONIŢĂ, A.; LEPĂDATU, C.:** Open Source Eminescu's Manuscripts: A Digitization Experiment, Studies in Informatics and Control, 19(1), 2010, pp. 79-84.
7. **DUMITRESCU, G.; LEPĂDATU, C.; CIUREA C.:** Creating Virtual Exhibitions for Educational and Cultural Development, INFOREC Publishing House, Informatica Economică Journal, 2014, 18(1), pp. 102-110.
8. **ENĂCHESCU, D.:** Data Mining: metode şi aplicaţii, Edit. Academiei Române, 2009, 277 p.
9. **FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P.:** From Data Mining to Knowledge Discovery in Databases, AAAI, AI Magazine, 17 (3), 1996, pp. 37-54.
10. **FILIP, F.-G.:** Decizie asistată de calculator: decizii, decidenţi - metode de bază şi instrumente informatice asociate, Ed. a 2-a, Bucureşti, Editura Tehnică, 2005, 376 p.
11. **FILIP, F.-G. HERERA-VIEDMA, E.:** Big Data in the European Union, National Academy of Engineering (NAE), SUA, Winter Bridge: A Global View of Big Data, 2014, 44(4), pp. 33-37.
12. **HAN, J.; KAMBER M.; PEI, J.:** Data Mining: Concepts and Techniques, Third Ed., Elsevier, 2011, 703 p.
13. **HASTIE, T.; TIBSHIRANI, R., FRIEDMAN, J.:** The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition, Springer-Verlag New York, 2009, 745 p.
14. **IONIŢĂ, A.; LEPĂDATU, C.; DUMITRESCU, G.:** Digital Cultural Landscape Content, Hernik, Jozef (ed.) Cultural Landscape – Across Disciplines, Oficyna Wydawnicza BRANTA, Kracow, Poland, 2009, pp. 255-277.
15. **LEPĂDATU, C.:** De la descriere bibliografică la web semantic, Academica, 2006, XVI (185-186/48-49), pp 78-81 şi XVI (188/51), pp. 42-85.
16. **LEPĂDATU, C.:** Support Systems for Knowledge Culture based on Solution and Tools from the Field of Business Intelligence – SSCBI, Proceedings of the Workshop IST – Multidisciplinary Approaches, Bucharest, Romania, 2006, pp. 7-12.
17. **LEPĂDATU, C.:** Acquisition Policy of a Library and Data Mining Techniques, Studies in informatics and control, 16(4), 2007, pp. 413-420.
18. **LEPĂDATU, C.:** Explorarea datelor şi descoperirea cunoştinţelor - probleme, obiective şi strategii, Revista Română de Informatică şi Automatică, 2012, 22(4), pp. 5-14.

19. **LEPĂDATU, C.:** Metode exploratorii multidimensionale, Revista Română de Informatică şi Automatică, 23(1), 2013, pp. 14-30.

20. **LEPĂDATU, C.:** Sisteme suport pentru decizii şi bibliomining, Revista Română de Informatică şi Automatică, 24(2), 2014, pp. 17-30.

21. **LEPĂDATU, C.:** Sisteme suport pentru decizii de bibliotecă, Revista Română de Informatică şi Automatică, 24(3), 2014, pp. 5-17.

22. **MAIMON, O. ROKACH, L. (EDS.):** Data Mining and Knowledge Discovery Handbook, 2nd Ed., Springer New York Dordrecht Heidelberg London, 2010, 1306 p.

23. **NICULESCU, C.; LEPĂDATU, C.; ŞTEFĂNESCU, D.:** SSCBI - A Teleworking Environment of Support Systems for Knowledge Culture. In the CD REV 2007 Proceedings of the International Conference Remote Engeneering Virtual Instrumentation, Porto, Portugal, iunie 2007.

24. **TUFFÉRY, S.:** Modélisation Predictive et Apprentissage Statistique avec R, TECHNIP, 2015, 415 p.

25. **VAPNIK, V. N.:** Statistical learning theory, Wiley-Interscience, 1998, 768 p.