METODE DE RECUNOAŞTERE A FORMELOR ÎN ANALIZA ECONOMICO – FINANCIARĂ

Prep.Drd. Eugeniu Tudor
Academia de Studii Economice,Bucureşti, Piaţa Romană nr.6, sect.1
teugeniu2003@yahoo.com

Abstract : Această lucrare are doua obiective : să prezinte metodele de recunoaştere supervizată utilizând tehnicile de tip kernel si implementarea acestor tehnici pentru clasa societatilor tranzactionate pe bursă. In prima secţiune se prezintă principiul general al sistemelor de recunoaştere supervizată precum si condiţiile care trebuiesc indeplinite pentru a asigura o generalizare „bună" a acestora. In a doua secţiune se prezintă problema generală a clasificării precum şi determinarea clasificatorilor in cazul utilizarii hiperplanelor optimale de separare. A treia sectiune prezinta implementarea „support vector machines" pentru piaţa bursieră din România. Ca set de formare au fost considerate societăţile care sunt tranzacţionate pe Bursa de Valori Bucureşti cu rezultatele din a doua jumătate a anului 2008.

Cuvinte cheie**:** recunoaştere supervizată, clasificator linear, set de formare, risc structural, optimizare convexă.

În cadrul activităţii umane apare necesitatea existenţei unor metode realiste de soluţionare a problemelor, în condiţiile unei informaţii apriorice incomplete asupra proceselor şi fenomenelor, în multe domenii, cum ar fi : biologia, arheologia, meteorologia, criminalistica, psihologia, stiinţa politică, domeniul militar, marketing, analiza economico-financiară, asigurările. Aceasta revine la determinarea legii care guvernează sistemului respectiv, altfel spus a mecanismului care generează diversele etape, stari ale evoluţiei sistemului analizat. În general determinarea legii de conducere optimală poate fi facută numai atunci când caracteristicile sistemului, procesului sau fenomenului analizat sunt complet cunoscute. Dar în evoluţia sa în mediul în care se situează, altfel spus în mediul real, aceste caracteristici nu sunt cunoscute, de aceea o metodă de optimizare constă în obţinerea datelor privind comportarea procesului din masurătorilor „reale" efectuate în timpul desfăşurarii acestuia, ca sursă de informaţii pentru adoptarea deciziei optime. În urma analizei masurătorilor efectuate, legea de evoluţie determinată se va apropia în cele din urmă de cea dorită, iar performanţa întregului sistem se va situa pe o traiectorie asimptotic apropiată de cea „ideală", îmbunătăţindu-se treptat preformanţele sistemului analizat. Procesul prin care legea de coducere optimală a mecanismului aleator care guvernează evoluţia sistemului, procesului respectiv, este determinată utilizand informaţia obţinută în timpul funcţionării sistemului a cărui performanţă, eficienţa generală se îmbunatăţeşte în consecinţă e numit în mod curent proces de învăţare.

Însă pentru ca relaţia funcţionala astfel determinată să fie utilă este necesar să se realizeze o simplificare a realităţii, altfel spus, apare necesitatea de a determina anumite categorii sau clase cu o delimitare clară şi naturală, care sunt determinate pentru entităţile informaţionale considerate. Aceste clase trebuie să fie într-o relaţie biunivocă cu realitatea surprinsă, studiată în sensul că trebuie să aibe o seminificatie clară, concisă şi să ofere un anumit grad de generalitate precum şi o înterpretabilitate simplă şi naturală. Diferenţierea obiectelor pe categorii se face în funcţie de proprietăţile fundamentale ale obiectelor, iar criteriile de asociere sub formă de clase au la bază gradul de asemănare a proprietăţilor respectivelor obiecte, masurat în funcţie de magnitudinea valorilor acestor proprietăti. Problema de învatare poate fi privită ca o problemă de

estimare sau aproximare succesivă a valorilor necunoscute ale unei funcționale, care a fost aleasă de proiectant să caracterizeze fenomenul supus analizei.

Oamenii dispun de simturile naturale cum ar fi : văzul, auzul, pipaitul, mirosul; care le permit sa perceapă anumite proprietăți ale obiectelor pe care le pot analiza astfel şi să poată structura, ierarhiza sau clașifica aceste obiecte sub forma unor submulțimi specifice şi disjuncte. Există, însa, numeroase situatii în care simturile naturale ale indivizilor şi informatiile de care aceştia dispun nu mai sunt suficiente pentru a segmenta anumite mulțimi de obiecte sau a clasifica corect aceste obiecte. Aceste situatii apar în cazul obiectelor multidimensionale, adică în cazul obiectelor caracterizate de mai multe caracteristici, iar numărul obiectelor care trebuiesc clasificate este foarte mare. În această situație diferențierea obiectelor pe categorii specifice nu se mai poate face intuitiv, exclusiv pe seama simturilor naturale fiind necesar, în această situație să se apeleze la o serie de metode şi tehnici specifice de mare complexitate şi cu un solid fundament statistico-matematic.

Având la bază teoria statistică a învățării dezvoltată de către cercetătorii V. Vapnik şi A. Chervonenkis pe parcursul a aproximativ 30 de ani, clasificatorii liniari cu „margine optimă de separare" au fost utilizați în teoria recunoşterii formelor pentru recunoaşterea carcterelor alfabetice (Cortes şi Vapnik, 1995; Scholopf, Bruges şi Vapnik1996; Bruges şi Scholopf, 1997), identificarea diverselor obiecte (Blanz, Scholopf, Bulthoff, Vapnik şi Vetter, 1996), recunoaştere vocală (Schmidt, 1996). În econometrie acest tip de clasificator a fost în problema estimării parametrilor asociați unor modele de serii de timp (Muller, 1997; Mukherjee, Osuna şi Girosi, 1997) conducând uneori sau la aceleaşi rezultate sau la rezultate mult mai bune decât modelele „clasice". De asemenea clasificatorii liniari cu margine optimă de separare au fost folosiți cu succes în estimarea densităților de probabilitate (Weston, 1997) şi în analiza discriminantă (Stitson, 1997). În abordările ulterioare s-au efectuat generalizări ale ideii de bază a clasificatorilor din această familie, în funcție de separanilitatea claselor (Smola, Scholopf şi Muller, 1998), s-a stabilit legătura cu teoria regularizării (Girossi, 1998). Pe baza acestor abordări se consideră că un clasificator determinat asigură o capacitate mare de clasificare dacă se asigură un compromis între viteza de recunoaştere[1] şi abilitatea predictivă[2] a acestuia. În algoritmii de clasificare dezvoltați pe baza principiului regularizării (ASVM, SOR) față de situația inițială au fost introduse următorele modificări :

1) Distanța între planele de separare se determină prin identificarea simultană atât a vectorului normal asociat acestuia precum şi a coeficientului de translație față de originea spațiului caracteristicilor pe baza căruia se determină (Mangasarian şi Musicant, 2000);
2) Eroarea în cazul determinării hiperplanelor de separare nestrictă a fost evaluată in sensul metricii L_2 spre deosebire de utilizarea clasică a normei L_1 (Lee şi Mangasarian, 2001)

Deoarece se face pe baza spațiului cauzal inițial se poate întâmpla ca formele să nu fie liniar separabile atunci se utilizează transformarea spațiului inițial cu ajutorul funcțiilor nucleu[3] asistând in ultimul deceniu la adaptarea algoritmilor existenți în funcție de aceste transformări (Akaho, 2001; Hamerling, 2001; Girolami, 2001; Kuss şi Graepel, 2004).

[1] În mod uzual prin viteză de recunoaştere se întelege acea fracțiune din formele din setul de formare clasificate corect de clasificatorul determinat.

[2] Abilitatea predictivă reprezintă numărul de obiecte din setul de predicție, a căror apartență la clase este presupusă necuoscută, sunt clasificate corect de către clasificator.

[3] Sunt denumite şi funcții kernel.

Indiferent de abordarea utilizată, găsirea hiperplanului de separare este echivalentă cu rezolvarea unei probleme de programare pătratică, aceasta înseamnă că soluția găsită este globală, iar dacă aceasta nu este unică mulțimea soluțiilor globale este convexă, iar dacă funcția obiectiv este strict convexă atunci soluția este în general unică.

PROBLEMA ÎNVĂȚĂRII STATISTICE

Scopul oricărui clasificator este să determine funcția de decizie

$$f: X \to Y$$

unde : X este setul de formare, adică mulțimea obiectelor de clasificat numit si spațiul intrărilor;

Y este mulțimea etichetelor corespunzătoare claselor in care se împart obiectele din populația supusă analizei numit de regulă si spațiul ieşirilor.

Inițial a fost considerat cazul clasificării binare, iar in această situație mulțimea claselor Y este redată de mulțimea $\{-1, +1\}$. În cadrul proceselor de învătare supervizată mulțimea Y poate fi considerată ca o mulțime indicator corespunzătoare obiectelor clasificate astfel : calsificatorul va asocia $+1$ sau -1 după cum un obiect se va încadra sau nu în clasa din care face parte in mod natural. Pornind de la clasificarea binară, echivalentă cu separarea spațiului intrărilor in două regiuni s–a efectuat trecerea la cazul clasificării multicategoriale.

În cadrul formulării generale problemei de învătare supervizată se consideră că vectorii x din spatiul intrărilor X sunt genrați in R^n, $n < \infty$, independent după o lege de probabilitate multidimensională stabilită dar necunoscută $P(x)$. Un clasificator atribuie fiecărei forme de intrare x o valoare y, numită in cele mai multe cazuri ieşire dorită, după o distribuție de probabilitate $P(y|x)$ fixată, dar necunoscută. Setul de formare sau de antrenare este dat de eşantionul, de volum m de forma :

$$S = \{(x_1, y_1), (x_2, y_2), \dots (x_m, y_m)\}.$$

Obeservațiile din eşantionul S sunt considerate independente si identic repartizate dupa legea de probabilitate $P(x, y) = P(x) * P(y|x)$ asociată variabilei aleatoare bidimensională (X, Y). Se consideră un sistem de recunoştere supervizată capabil să genereze o fmilie de funcții $F_1(= \{f(x, "(")" (" "(" \in "(" \}$, unde Λ este mulțimea parametrilor necunoscuți. În acest fel problema învățării artificiale este aceea a aproximării funcției $f(x, ()$ care realizează cea mai bună aproximare $\check{y}$ a răspunsului y generat de sistem.

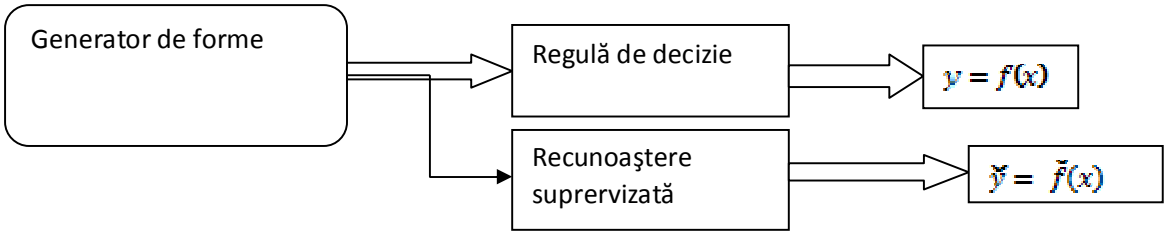


FIGURA 1 REPREZENTAREA PROBLEMEI GENERALE DE RECUNOŞTERE SUPERVIZATĂ. DUPĂ GENERAREA FORMELOR DE MECANISMUL ALEATOR ŞI SUNT ÎNCADRATE ÎN CLASELE ÎN CARE SE STRUCTUREZĂ POPULAȚIA INIȚIALĂ SE APROXIMEAZĂ REGULA DE CLASIFICARE PRIN $\check{y}$

De regulă, cea mai bună aproximare este determinată în sensul minimizării unei funcții care măsoară neconcordanța între y şi $\breve{y}$, numită funcție de eroare a calsificării şi notată cu $D(y,\breve{y})$, adică determinarea minimului funcționalei care cuantifică riscul :

$$R(\text{O} = \int\!\!\!\int D(y, f(x, \text{"}(\text{"}\ )\ ) dP(x,y)$$

Definiție : Funcția de eroare este funcția $c : XxXxY \rightarrow [0,\infty]$ *astfel încât* $c(x,y,y) = 0$ *. unde* $(x,y,f(x)) \in XxXxY$ *, iar* x *este o formă din* X *,* y *este clasa reală din care aceasta face parte si* $f(x)$ *este clasa în care este încadrată de clasificator.*

În general în problema clasicării se utilizează funcția eroare de forma următoare :

$$D(x,y,f(x)) = \frac{1}{2}|f(x) - y| \quad (2.1)$$

Dacă forma x din spațiul formelor este clasificată corect $z = D(x,y,\breve{y}) = 0,$ pe când în situția în care forma este clasificată greşit $z = D(x,y,\breve{y}) = 1$. Dacă se consideră funcția de eroare pentru întregul eşantion ea poate fi considerată o variabilă aleatoare. În această situație media acestei variabile aleatoare determinată pe toată distribuția $P(x,y)$ constituie funcționala care exprimă riscul asociat procesului de clasificare :

$$R(f) = \int_{XxY} \frac{1}{2}|f(x) - y| dP(x,y) \quad (2.2)$$

În cadrul metodelor de clasificare, riscul asociat acestui proces trebuie să fie, adică să se determine acea funcție f care respectă :

$$f_{opt} = \arg min_f (R[f]) \quad (2.3)$$

Expresia din relatia (2.3) reprezintă tocmai principiul riscului minim din teoria fiabilității.

Având în vedere ipotezele formulate in cazul descrierii problemei de învățare de la începutul acestei secțiuni se poate constata că determinarea funcției f_{opt} este dificilă dacă nu imposibilă în condițiile informației apriorice necunoscute. De aceea pentru a înlătura acest neajuns se consideră principiul minimizării riscului empiric[4]. Deoarece pentru solutionarea problemei determinării suprfețelor de separare se utilizează un eşantion de volum m atunci pentru a aproxima riscul se utilizează riscul empiric claculat după următoarea formulă :

$$R_{\text{emp}}(f_{\downarrow}(\ ) = 1/m\ \Sigma_{\text{i}}(l=1)^{\uparrow} m \| f(x_{\downarrow} l\ ) - y_{\text{i}} l\ | \quad (2.4)$$

unde $f_{\text{i}}($ este funcția din $F_{\text{i}}($ cu parametrul fixat λ.

Procedând ca in cazul determinării funcției f_{opt} din rel. (2.3) se va determina acea funcție $f_{\text{i}} opt = f_{\text{i}}((^{\uparrow} *)$ astfel încât :

$$(^{\uparrow} *= \arg\ [\![[\![min]\!]_{\text{i}} f\ (R_{\downarrow} emp\ (f_{\text{i}}(\ )\ )]\!] \quad (2.5)$$

Acest criteriu reprezentat de relația (2.5) este numit criteriul minimizării riscului empiric.

Totuşi utilizând acest criteriu în forma descrisă mai sus se poate întâmpla ca pe setul de formare clasificatorul determinat să asigure o clasificare corectă totală, insă să aibe o abilitate predictivă redusă.

DEDUCEREA VECTORILOR SUPORT

Ipoteza fundamentală pentru determinarea hiperplanelor de clasificare cu margine optimă este aceea că în spațiul variabilelor obiectele din populația inişială sunt liniar separabile. De

[4] Notat prescurtat ERM, de la Empiric Risk Minimisation

aceea în condițiile unei astfel de ipoteze obiectivul principal este acela al găsirii uneo funcții clasificator care va separa calsele şi va asigura în acelaşi timp distanța maximă între cele două clase pe care le separă. În condițiile în care populația inițială este liniar separabilă atunci pentru a împărți obiectele pe clase se pot determina mai multe hiperplane de separare, astfel prin hiperplanul optim de separare se determină „cel mai bun" hiperplan de separare din mulțimea hiperplanelor admisibile. În aceste condiții hiperplanul optim de separare va trece prin „mijlocul" punctelor din cele două clase pe care le separă, fiind considerat hiperplanul cel mai sigur deorece dacă se consideră un individ neclasificat sau care nu este descris complet o mică eroare în caracterizarea sa nu va modifica procesul de clasificare dacă distanța față de hiperplan este mare asigurându-se astfel robustețea claselor determinate. Clasificatorul cel mai bun este cel care asigură marginea cea mai mare între formele din setul de antrenare. De asemenea determinarea marginii maxime dintre clase are în mod natural riscul clasificării unei forme necunoscute mai mic pe baza supoziției că minimizarea numărului de erori pe setul de formare asigură un nivel de discriminare ridicat între formele neclasificate.

Fie $S = \{(x_1, y_1), (x_2, y_2), \dots (x_m, y_m)\}$ setul de formare unde x_i este forma asociată obiectului „i" , $i = \overline{1,m}$, iar y_i este clasa asociată obiectului „i", care se presupune cunoscută, $i = \overline{1,m}$. Pentru determinarea clasificatorului cu margine optimă de clasificare se va considera problema clasificării binare, corespunzătoare situației în care $y_i = \{\pm 1\}$. Clasificatorul este considerat liniar dacă funcția de decizie asociată acestuia poate fi descrisă printr-o funcție liniară de variabilele care o identifică o formă :

$$f(x) = w^t * x + b \quad (1)$$

unde $w \in R^n, b \in R$ fiind necunoscuți pot fi considerați parametrii modelului.

Deoerece hiperplanul considerat împarte din punct de vedere geometric spațiul R^n în două regiuni, asociate cu cele două clase atunci pentru a clasifica o formă necunoscută este suficient să se considere funcția semn asociată acesteia, adică :

$$y = sgn(f(x))$$

Fie (x_i, y_i) o formă astfel încât $y_i = 1$ şi (x_j, y_j) o altă formă aszfel încât $y_j = 1$. Deoarece $y_i = 1$ atunci $f(x_i) \geq 0, adică\ w^t * x_i + b \geq 0,$ aceasta fiind echivalentă cu $y_i(w^t * x_i + b) \geq 0,$ pentru fiecare obiect din prima clasă. Pe baza aceluiaşi raționament pentru cazul al doilea se obține : $y_j(w^t * x_j + b) \geq 0$ pentru fiecare obiect din cea de-a doua clasă.

Definiție : Setul de formare S este liniar separabil dacă şi numai dacă $\exists\, w \in R^n, b \in R$ *astfel încât :* $y_i(w^t * x_i + b) \geq 0,\ i = \overline{1,m}$ *(2).*

În situația în care o formă este clasificată eronat, de exemplu $y_i = 1$ cu toate că $w^t t * x_i i + b) \leq 0$ atunci expresia (2) este negativă şi poate fi folosită ca un termen de penalizare în cadrul algoritmilor de determinare a hiperpanului optim de separare.

Distanța de la un obiect x_i la panul de separare dat de $w^t * x + b$ este dată de $d(x_i) = |w^t t * x_i i + b| / \|w\|$ sau $d(x_i) = y_i \left(\frac{w^t * x_i}{\|w\|} + \frac{b}{\|w\|}\right)$ (3) unde $y_i = \pm 1$.

Dacă un obiect este clasificat eronat atunci $d(x_i) < 0$, semnul minus asociart expresiei arată că forma încadrată greşit se găseşte în partea opusă a clasei din care face parte.

Pentru hiperplanul dat de (1) se pot considera două forme din clase diferite astfel încât

$$f(x_1) = 1 \quad (4)$$
$$f(x_2) = -1 \quad (5)$$

Din relațiile (4) și (5) se obține sistemul :

$$\begin{cases} w^t * x_1 + b = 1 \\ w^t * x_2 + b = -1 \end{cases}$$

Scazând ecuațiile din sistemul de mai sus se obține :

$$w^t(x_1 - x_2) = 2$$

Care poate fi scrisă în forma echivalentă :

$$\frac{w^t(x_1 - x_2)}{\|w\|} = \frac{2}{\|w\|}$$

Deoarece $d(x_1) = \frac{1}{\|w\|}$ și $d(x_2) = \frac{1}{\|w\|}$ atunci pentru a detremina marginea maximă între clase trebuie să minimizăm norma vectorului w. Hiperplanele date de (4) și (5) sunt numite hiperplane canonice. Astfel problema determinării hiperpalnelor cu margine optimă devine :

$$P_1 \begin{cases} min_w \frac{1}{2} \|w\| \\ y_i(w^t * x_i + b) \geq 1, i = \overline{1,m} \end{cases}$$

$$P_2 \begin{cases} min_w \frac{1}{2} \|w\|^2 \\ y_i(w^t * x_i + b) \geq 1, i = \overline{1,m} \end{cases}$$

Fie . Atunci soluția problemei P_1 este aceeași cu a problemei P_2.

Demonstrație :

Presupunem că w_1 este soluția problemei P_1 și w_2 este soluția problemei P_2 și în plus. Dacă w_1 este soluția lui P_1 atunci restricțiile sunt satisfăcute, adică este o soluție fezabilă și ținând seama de modul de definiție al soluției avem :

$$\frac{1}{2} \|w_2\| \geq \frac{1}{2} \|w_1\|$$

De unde se obține după simplificare : $\|w_2\| \geq \|w_1\|$ (6)

Dacă w_2 este soluția problemei P_2 atunci restricțiile asociate acesteia sunt satisfăcute și avem :

$$\frac{1}{2} \|w_1\|^2 \geq \frac{1}{2} \|w_2\|^2$$

După simplificare se obține : $\|w_1\|^2 \geq \|w_2\|^2$ (7)

Din (6) și (7) se obține că : $\|w_1\|^2 = \|w_2\|^2$ adică cele două probleme au aceeași valoare optimă. Prin urmare pentru determinarea marginii maxime dintre clase se poate rezolva problema P_2. Fie $L(w, b, \alpha)$ funcția lui Lagrange asociată problemei P_2. Atunci avem :

$L(w, b, \alpha) = 1/2 \|w\|^2 - \sum_i (i = 1)^m \alpha_i [y_i (w^T * x_i + b) - 1]$, unde α_i sunt multiplicatorii lui Lagrange . În funcție de restricțiile cărora le sunt asociați multiplicatorii lui Lagrange indeplinesc urmaătoarele condiții :

1) Dacă restricția este de tipul $r_i(x) \geq 0$ atunci $\alpha_i \geq 0$
2) Dacă restricția este de tipul $r_i(x) = 0$ atunci $\alpha_i \in R$
3) Dacă restricția este de tipul $r_i(x) \leq 0$ atunci $\alpha_i \leq 0$

Pentru cazul problemei de optimizare cu restricții de mai sus atunci avem $\alpha_i \geq 0, i = 1,2, \dots, m$

$$\Delta L/\Delta w\ (w, b, "(")\ = w - \Sigma_{\downarrow}(i=1)^{\uparrow} m \equiv [(_{\downarrow} i\ y_{\downarrow} i\ x_{\downarrow} i\ ]\ = 0\ (8)$$

$$\Delta L/\Delta b\ (w, b, "(")\ = \Sigma_{\downarrow}(i=1)^{\uparrow} m \equiv [(_{\downarrow} i\ y_{\downarrow} i\ ]\ = 0\ (9)$$

$$\Delta L/\Delta(\ (w, b, "(")\ = 1 - y_{\downarrow} i\ (w^{\uparrow} t * x_{\downarrow} i + b\ ) \leq 0\ (10)$$

$$(_{\downarrow} i\ (1 - y_{\downarrow} i\ (w^{\uparrow} t * x_{\downarrow} i + b\ )\ ) = 0\ \ (11)$$

Din (8) se obține : $w = \Sigma_{\downarrow}(i=1)^{\uparrow} m \equiv [(_{\downarrow} i\ y_{\downarrow} i\ x_{\downarrow} i\ ]$ (12). Inlocuind în expresia funcției lui Lagrange obținem :

$L(w, b, "(")\ =\ \ 1/2\ \ \Sigma_{\downarrow}(i, j=1)^{\uparrow} m \equiv [(_{\downarrow} i\ (_{\downarrow} j\ ]\ \ y_{\downarrow} i\ y_{\downarrow} j\ x_{\downarrow} i^{\uparrow} t\ x_{\downarrow} j - \Sigma_{\downarrow}(i, j=1)^{\uparrow} m \equiv [(_{\downarrow} i\ (_{\downarrow} j\ ]\ \ y_{\downarrow} i\ y_{\downarrow} j\ x_{\downarrow} i^{\uparrow} t\ x_{\downarrow}$

Pornind de la problema primală (P_2) se obține următoarea problemă duală în sensul lui Wolfe :

$$\{\max_{\top}((_{\downarrow} i)\ [\Sigma_{\downarrow}(i=1)^{\uparrow} m \equiv (_{\downarrow} i\ \ - 1/2\ \ \Sigma_{\downarrow}(i, j=1)^{\uparrow} m \equiv [(_{\downarrow} i\ (_{\downarrow} j\ ]\ \ y_{\downarrow} i\ y_{\downarrow} j\ x_{\downarrow} i^{\uparrow} t\ x_{\downarrow} j\ ]\ @\Sigma_{\downarrow}(i=1)^{\uparrow} m \equiv [(_{\downarrow} i\ x$$

Prin rezolvarea dualei se determină vectorul w. Vectorii suport corespund multiplicatorilor nenuli şi se găsesc pe hiperplanele canonice. Pentru a detremina scalaru b se folosec ecuațiile vectorilor suport din problema primală asociată pentru vectorul w determinat din problema duală.astfel dacă setul de formare se reduce numai la vectorii suport atunci hiprplanul determinat este acelaşi ca în situația în care setul de formare era constituit din mai multe forme.

Dacă in spațiul variabilelor clasele nu sunt separabile liniar atunci în problema primală în cadrul expreisiei de minimizat apare un termen de penalizare, problema rescriindu-se astfel [4]:

$$min_w \frac{1}{2}\ \|w\|^2 + C \sum_{i=1}^{m} \varepsilon_i$$

$$y_i(w^t * x_i + b\ ) \geq 1, i = \overline{1, m}$$

a cărei dual in sensul lui Wolfe este :

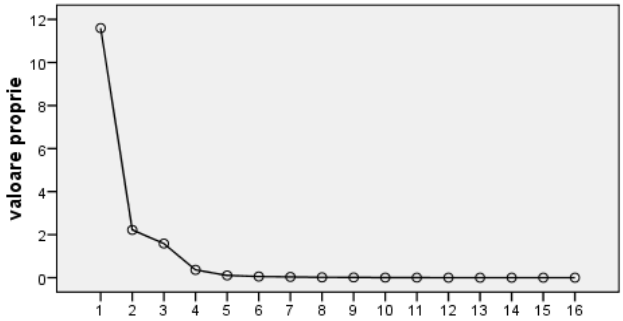
$$\{\max_{\top}((_{\downarrow} i)\ [\Sigma_{\downarrow}(i=1)^{\uparrow} m \equiv (_{\downarrow} i\ \ - 1/2\ \ \Sigma_{\downarrow}(i, j=1)^{\uparrow} m \equiv [(_{\downarrow} i\ (_{\downarrow} j\ ]\ \ y_{\downarrow} i\ y_{\downarrow} j\ x_{\downarrow} i^{\uparrow} t\ x_{\downarrow} j\ ]\ @\Sigma_{\downarrow}(i=1)^{\uparrow} m \equiv [(_{\downarrow} i\ x$$

În formularea problemei inițiale cazul în care populația nu este liniar separabilă apare teremenul de penalizare ε care modifică marginea dintre clase, conducând la situația în care anumite forme sunt clasificate eronat.

DETREMINAREA VECTORILOR SUPORT PENTRU SOCIETĂȚILE TRANZACȚIONATE PE BURSĂ

Ca aplicație a celor descrise mai sus se consideră un eşantion format din 74 de firme tranzacționate pe Bursa de Valori Bucureşti,observațiile necesare efectuării au fost obținute din situațiile financiare depuse de către acestea pe trimestrul trei al anului 2008. Astfel au fost considerate ca variabile : total activele imobilizate, total active circulante, datorii pe termen scurt, datorii pe termen lung, venituri în avans, capital social subscris şi vărsat, creanțe, cifra de afaceri netă, venituri din exploatare, cheltuieli din exploatare, venituri financiare, cheltuieli financiare, furnizori restanți, plăți restante şi numarul de angajați ai fiecărei firme. În perioada considerată este identificată o scădere a tranzacțiilor bursiere, iar acest lucru poate fi pus în legătură şi cu

creşterea riscului având în vedere eficiența scazută înregistrată de către firmele cotate. În aceste condiții se consideră că fiecare firmă este identificată prin vectorul corespunzător format din cele $16\ variabile$, altfel spus prin vectorul $16-dimensional$ asociat. Pentru a reduce dimensionalitatea se foloseşte transformarea spațiului inițial prin utilizarea analizei în componente principale. Astfel prin aplicarea acestei tehnici se obțin trei noi variabile, reținute conform criteriului lui Kaiser, care explică 96,27% din variabilitatea inițială. De asemenea variabilele considerate sunt bine „alese", deoarece gradul de suprapunere informațională între variabilele inițiale şi variabilele noi este mare, din fiecare variabilă recuperându – se peste 90% din variabilitatea sa.



După cum se observă din reprezentarea valorilor proprii ale matricei de corelație asociată variabilelor inițiale, care nu reprezintă altceva decât varianța noilor variabile, pentru a reprezenta spațiul inițial sunt necesare trei noi variabile corespunzătoare valorilor proprii mai mari decât unu. Se realizează in acest mod sinteza informațională trecându-se de la reprezentarea inițială la una în care fiecare societate considerată este identificată prin cele trei noi variabile. Din legătura cu spațiul inițial se poate considera că, în spațiul transformat, prima variabilă corespunde unui indicator de eficiența generală, a doua variabilă fiind puternic „corelată" cu variabilele furnizori restanți şi plăți restante poate fi considerat ca un indicator privind datoriile istorice ale firmelor, iar cea de-a treia variabilă fiind influențată de variabilele datorii pe termen lung şi venituri în avans este privită ca un indicator de dezvoltare pentru firma respectivă.

Cele trei variabile „compuse" obținute sunt necorelate, fiind asociate unor axe ortogonale, altfel spus ele sunt necorelate. Pe baza celor trei variabile identificate după reducerea dimensionalității se trece la segmentarea spațiului inițial. Segmentarea presupune obținerea a două –trei clase corespunzătoare unor situații reale concrete. Astfel pentru setul de firme considerate se obține două clase reprezentative : prima clasă conține firmele care sunt caracterizate printr-o eficiență scazută înregistrând pentru variabila compusă eficineță valorile cele mai mici, de asemenea sunt considerate după cea de-a treia firme lipsite de perspectivă, iar după a doua variabilă sunt firme care intîmpină dificultăți din perspectiva datoriilor. Spre deosebire de firmele din prima clasă cele din clasa a doua sunt caracterizate printr-o eficiență mai ridicată, dar pot fi considerate şi firme care încearcă să asigure un echilibru intre datoriile pe care le au şi veniturile realizate.

În urma divizării spațiului inițial prima clasă conține 29 de firme iar cea de-a doua clasă este formată din 20 de firme. Pentru aceste două clase se determină hiperplanul optim de separe folosind abordarea pe baza funcției lui Lagrange. Deoarece problema determinării vectorului

normal este o problemă de programare pătratică de regulă soluția este determinată cu aproximație pe baza considerațiilor legate de inversarea matricilor de dimensiune mare.

Astfel pentru setul de formare, conținând 49 de firme s-a folosit pentru determinarea sa în procesul iterariv o eroare asociată de 10^{-5} , iar numărul iterațiilor a fost de ales la 1000. În consecință pentru setul de formare s-a obținut următoarele valori pentru vectorul normal asociat : (-5.9426, 6.2227, -2.4193), iar termenul liber al funcției de clasificare este 0.5736, deci ecuația hiperplanului de separare este :

$$f(x) = -5.9426 * Eficienta + 6.2227 * Datorii\ Istorice - 2.9426 * Dezvoltare + 0.5736$$

iar marginea asociată acestuia este de 0. 0092, obținîndu-se o acuratețe de 93.88%, altfel spus procentul de forme clasificate corect. Pentru a valida acest clasificator restul formelor au fost grupate în setul de predicție. Prin folosirea regulii de decizie dată de apartenența unui obiect la o clasă sau alta se face după semnul funcției discriminant determinată se constată ca pentru setul de predicție procentul de forme clasificate corect este de 96.37% ceea ce arată că pentru eşantionul de firme selectate acesta asigură discriminarea corectă a acestora.

După cum se poate constata pentru firmele considerate hiperpalnul de separare asigură o segregare corectă a firmelor, insă se poate constata că pentru trimestrul trei al anului 2008 activitatea firmelor este afectată fapt pus în corelație cu variabila numită generic „Eficiența", care având varianța cea mai mare înregistrează valori reduse pentru firmele considerate. De asemenenea ca firme eficiente sunt în general firmele care activiează în industria petrochimică şi firmele care furnizează energie electrică, insă acestea sunt considerate domenii privilegiate în comparație cu celelalte. Pentru setul de antrenare aceste firme au fost st eliminate şi au fost considerate în setul de predicție.

Pentru validarea hiperplanului de separare s-au efectuat 1000 reselecții pentru formele din setul de formare eliminându-se un număr de 10 forme alese la întâmplare, iar pentru reselecțiile efectuate s-a obținut o medie a numărului de forme clasificate corect de 92%, procesul fiind afectat eliminarea în cadrul unor selecții de eliminarea firmelor suport considerate.

Anexa nr.1

	Initial	Extraction
activeimobtot	1	0.97
activecirctot	1	0.991
datoriiscurt	1	0.989
datoriicurentenete	1	0.929
datoriipetermenlung	1	0.99
venituriinavans	1	0.968

capsocsubscrissivarsat	1	0.991
creante	1	0.958
cifradeafacerineta	1	0.996
vendinexploatare	1	0.997
cheltdinexploatare	1	0.993
venfin	1	0.988
cheltfin	1	0.995
furnizorirestanti	1	0.907
platirestante	1	0.863
nrdeangajati	1	0.88

Tabelul nr.1 Variabilitatea explicată de spațiul redus

		Component	
	1	2	3
activeimobtot	0.948	-0.074	0.255
activecirctot	0.953	0.184	0.222
datoriiscurt	0.884	0.433	0.14
datoriicurentenete	0.649	-0.625	0.343
datoriipetermenlung	-0.002	0.169	0.98
venituriinavans	0.43	-0.139	0.874
capsocsubscrissivarsat	0.984	0.076	0.13
creante	0.87	0.368	0.255
cifradeafacerineta	0.956	0.229	0.171
vendinexploatare	0.959	0.221	0.168
cheltdinexploatare	0.936	0.295	0.171
venfin	0.993	0.036	0.013
cheltfin	0.967	0.244	0.009
furnizorirestanti	0.514	0.801	-0.036
platirestante	0.165	0.903	0.14
nrdeangajati	0.919	-0.103	0.155

Tabelul nr.2 Importanța variabilelor inițiale în cadrul variabilelor compuse

Bibliografie :

1. J.A. Hartigan . *Clustering Algorithms*. John Wiley 1975

2. G. Ruxanda . *Analiza Datelor*. Ed ASE Bucuresti 2001

3. B.E. Boser, I.M. Guyon, V. Vapnik. *A Training Algorithm for Optimal Margin Classifiers*.Fifth annual workshop on computational learning theory 144-152 Pittsburg 1992

4. B. Scholkopf , A. J. Smola . *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2002

5. E.L. Allwein, R.E. Schapire,Y. Singer. *Reducing multiclass to binary: A unifying approach for margin classifiers*. 17th International Conf. Machine Learning (P. Langley, ed.) 9–16. Morgan Kaufmann, San Francisco, 2000

6. F.R. Bach, M.I. Jordan. *Kernel independent component analysis*. Journal of Machine Learning 2002.

7. I. Steinwart. *Support vector machines are universally consistent*. *J. Complexity* 18 , 2002

8. V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Berlin , 1982.

9. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995

10. V. Vapnik . *Statistical Learning Theory*. Wiley, New York, 1998

11. V. Vapnik, A. Chervonenkis . *The necessary and sufficient conditions for consistency in the empirical risk minimization method*. *Pattern Recognition and Image Analysis* 1 283–305, 1991.

12. J.Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, New York, 1990.