Cuantificarea dependentei spatiale în clasificarea automata

Conf.dr. Emil STAN
Academia de Politie “Alexandru Ioan Cuza”, Bucuresti

În cadrul articolului se analizeaza problema cuantificarii dependentei spatiale care poate fi identificata cu reducerea numarului de elemente din fiecare multime de valori posibile ale fiecarui criteriu de clasificare. Reducerea trebuie sa pastreze imaginea spatiala a secventei de date pentru a nu se afecta clasele obtinute de algoritm.
***Cuvinte cheie:** dependenta spatiala, cuantificare, clasificare automata, partitie, clase, criteriu de clasificare, recunoasterea formelor.*

Introducere

În prezent, o mare atentie se acorda algoritmilor de clasificare, datorata multitudinii domeniilor de aplicare.
Presupunem ca fiecare din obiectele (formele) analizate este descris cu parametrii {x1, x2, x3,…, xN}. Se introduce spatiul parametrilor G, în care fiecare obiect concret gi $=\{x^i_1, x^i_2, .., x^i_N\}$ reprezinta un punct $g_i \in G$.
Fie cîteva clase de obiecte (mai departe, pentru usurinta, presupunem ca sînt doua A si B). În domeniul recunoasterii formelor în algoritmi apar obiectele sirului director $g_1, g_2, g_3, \ldots g_N$ cu informatii asupra clasei la care acestea apartin [1]. În domeniul clasificarii algoritmii studiaza obiecte neidentificate si se cere algoritmilor sa le identifice concret [2].
Pentru rezolvarea acestei probleme, majoritatea algoritmilor conduce fie la construirea în spatiul G a unei suprafete care sa separe punctele multimii A de punctele multimii B, fie construirea unei functii de decizie, care minimizeaza, într-un sens dat, probabilitatea luarii unei decizii gresite.
Rezolvarea problemei nu este posibila daca nu se fac cîteva ipoteze asupra distributiei punctelor în spatiul G.
La baza multor algoritmi existenti sta ipoteza de lucru legata de formularea urmatoarei idei despre distribuirea reciproca a punctelor claselor A si B: daca doua puncte stau într-un anumit sens aproape unul de celalalt, atunci asa cum este si normal, ele apartin aceleiasi clase .
Daca acum vom presupune ca distributia punctelor în G este astfel încât clasele corespund unor puncte suficient de departate între ele, atunci problema construirii pentru recunoasterea formelor se poate transforma în problema distingerii unor astfel de clase si prin urmare, prin rezolvarea acestei probleme, în principiu, se poate evita necesitatea unor informatii suplimentare despre apartenenta fiecarui punct al sirului director la vreuna din clase .
Vom numi problema distingerii unor astfel de grupuri problema clasificarii automate (instruirea fara profesor, autoinstruire, cluster analysis, taxonomie, stratificatii, taximetrie, morfometrie, botriologie etc.).

Reducerea numarului de elemente din multimile de valori

Problema cuantificarii se identifica cu reducerea numarului de elemente în fiecare multime de valori astfel ca sa se pastreze cel mai bine imaginea spatiala a secventei de date. Expresia “cel mai bine“ este folosita în absenta unei notiuni riguroase pentru gruparea împreuna a tuturor nivelurilor “gri“ care sînt cel mai dependente spatial.
Fie Zx= {1,2,..,m} si Zy= { 1,2,..,n} domeniile spatiale x si y. Fiecare pereche $(z1,z2) \in Zx \times Zy$ defineste locul din spatiu din sirul imaginii de date I .
Fie Li=$\{\{l_{i1}, l_{i2}, .., l_{iNi}\}$ multimea de valori pentru criteriul Xi, i=1,2,..,N. Presupunem ca fiecare Li este ordonat crescator $l_{ij} < l_{ik}$ pentru j < k. Spatiul masuratorilor este $G = L_1 \times L_2 \times \ldots \times L_N$. Secventa imagi-

nii datelor I poate fi reprezentata ca functia: $I: Z_x \times Z_y \to G$, care poate fi scrisa direct ca o secventa bidimensionala:

$$I = \begin{pmatrix} g_{11} g_{12} \cdots g_{1n} \\ g_{21} g_{22} \cdots g_{2n} \\ . \\ . \\ g_{m1} g_{m2} \cdots g_{mn} \end{pmatrix}$$

Natura multimii imagine a lui I poate fi vazuta daca îl scriem ca N secvente separate. Fie

$$I_y = \begin{pmatrix} g^j_{11} g^j_{12} \cdots g^j_{1n} \\ g^j_{21} g^j_{22} \cdots g^j_{2n} \\ . \\ . \\ g^j_{m1} g^j_{m2} \cdots g^j_{mn} \end{pmatrix} \quad j=1,2,..,N$$

unde g^j_{rs} este a j-a componenta a elementului g_{rs} în secventa I, adica $g_{rs} = (g^1_{rs}, g^2_{rs}, .. g^N_{rs})$.

Fiecare $g^j_{rs} \in L_j$ apare ca o nuanta de gri (nici alba, nici neagra) în punctul de coordonate (r,s) al imaginii obtinuta din criteriul X. Definim pentru fiecare pereche de coordonate spatiale $(z_1, z_2) \in Z_X \times Z_Y$ si fiecare submultime A a lui $Z_X \times Z_Y$, $A \subset Z_X \times Z_Y$, distanta între coordonatele (z_1, z_2) si submultimea A prin :

$$\mathrm{dist}((z_1, z_2), A) = \min_{(y1,y2)\in A} d((z_1, z_2), (y_1, y_2)) \quad \text{unde}$$

$$d((z_1, z_2), (y_1, y_2)) = (z_1 - y_1)^2 + (z_2 - y_2)^2$$

Fie $\propto$ parametrul care fixeaza distanta de depasire, iar θ parametrul de prag al dependentei spatiale. Vom construi acum o partitie a lui L_j, j=1,2,..,N.

Matricea frecventelor relative

Pentru fiecare imagine I_j, dorim sa definim o matrice $P^j = (P^j_{ik})$ unde P^j_{ik} este frecventa relativa a imaginii a j-a în care doua puncte distincte de distanta maxima d unul fata de celalalt , au densitati de gri l_{ji} si l_{jk} respectiv.

$$P^j_{ik} = \frac{\mathrm{card}\{(z_1, z_2) \in Z_X \times Z_Y | 0 < \mathrm{dist}((z_1, z_2), I_j^{-1}(l_{ij}) \le \alpha \, ^o \text{ i } I(z_1, z_2) = l_{jk}\}}{\mathrm{card}\{Z_X \times Z_Y\}}$$

Observam ca P^j este o matrice simetrica din cauza ca urmatoarele doua propozitii sînt logic echivalente:

$$\exists (z_1, z_2): 0 < \mathrm{dist}((z_1, z_2), I_j^{-1}(l_{ij}) \le \alpha \, ^o \text{ i } I_j(z_1, z_2) = l_{jk}$$

$$\exists (z_3, z_4): 0 < \mathrm{dist}((z_3, z_4), I_j^{-1}(l_{ij}) \le \alpha \, ^o \text{ i } I_j(z_3, z_4) = l_{ji}$$

Aceasta se poate vedea fiindca prima propozitie este echivalenta cu:

$$\exists (z_1, z_2): 0 < \mathrm{dist}((z_1, z_2), I_j^{-1}(l_{ji}) \le \alpha) \, ^o \text{ i } I_j(z_1, z_2) = l_{jk}$$

$$\Leftrightarrow \exists (z_1, z_2): 0 < \min_{(y_1, y_2) \in I_j^{-1}(l_{ji})} d((z_1, z_2), (y_1, y_2)) \le \alpha \, ^o \text{ i } I_j(z_1, z_2) = l_{jk}$$

$$\Leftrightarrow \exists (z_1, z_2), \exists (z_3, z_4): 0 < d((z_1, z_2), (z_3, z_4) \le \alpha, I_j(z_3, z_4) = l_{ji} \, ^o \text{i}$$

$$I_j(z_1, z_2) = l_{jk} \Leftrightarrow \exists (z_3, z_4), \exists (z_1, z_2): 0 < d((z_3, z_4), (z_1, z_2) \le \alpha, I_j(z_1, z_2) = l_{jk}$$

$$^o \text{ i } I_j(z_3, z_4) = l_{ji} \Leftrightarrow \exists (z_3, z_4): 0 < \mathrm{dist}((z_3, z_4), I_j^{-1}(l_{jk}) \le \alpha \, ^o \text{ i } I_j(z_3, z_4) = l_{jk}$$

$$\text{Simetrizãm } P_{ij} = P_{ji} = \frac{P_{ij} + P_{ji}}{2}$$

P^j este matricea continînd informatia dependentei spatiale a densitatilor gri pentru punctele aflate în vecinatate. În continuare normalizam matricea P^j. Definim $Q^j = (q^j_{ki})$ matricea $N_i \times N_j$ data prin:

$$q_{ki}^{j} = \begin{cases} \dfrac{p_{ki}^{j}}{p_{k}^{j} p_{i}^{j}} & \text{cînd } p_{k}^{j}, p_{i}^{j} \neq 0 \\ 0 \text{ în celelalte cazuri} \end{cases}$$

unde $p_{k}^{j} = \sum_{i=1}^{N_J} p_{ki}^{j}$ º i $p_{i}^{j} = \sum_{k=1}^{N_j} p_{ki}^{j}$, q_{ki}^{j} , reflecta suma nivelelor gri ljk si lji care sunt relative la imaginea j. $q_{ki}^{j}=0$ înseamna ca nu sînt spatial apropiate, $g_{ki}^{j}<1$ înseamna ca sunt negativ spatial dependente (tonul gri ljk nu va apare alaturi de tonul gri lji) $q_{ki}^{j}=1$ înseamna ca sînt spatial independente si $q_{ki}^{j}>1$ înseamna ca sînt spatial dependente.

Constructia relatiei de echivalenta

Pentru o valoare data θ a pragului de dependenta spatiala putem defini relatia binara

$$R^{j}(\theta) \text{ pe } L_{j}, R^{j}(\theta) = \{(l_{jk}, l_{ji}) \in L_{j} \times L_{j} | q_{ki}^{j} > \theta\} \cup \{(l_{jk}, l_{jk}) | 1 \leq k \leq N_{j}\}$$

Deoarece P^{j} este simetrica, Q^{j} este de asemenea simetrica si relatia $R^{j}(\theta)$ este simetrica. Regularitatea spatiala naturala a imaginii garanteaza ca dependenta spatiala a punctelor aflate în vecinatate este maxima pentru acelasi nivel de gri. Deci $R^{j}(\theta)$ este reflexiva. Mai avem nevoie sa facem $R^{j}(\theta)$ tranzitiva pentru a o face relatie de echivalenta. Pentru aceasta calculam închiderea tranzitiva $R^{j}(\theta)^{T}$ a lui $R^{j}(\theta)$.

$$R^{j}(\theta)^{T} = \bigcup_{n=1}^{N_j} \underbrace{R^{j}(\theta) o R^{j}(\theta) o \ldots R^{j}(\theta)}_{\text{de n ori}} \quad \text{unde}$$

"o" este operatia binara a compunerii relatiilor.

Desi $R^{j}(\theta)^{T}$ este cea mai mica relatie tranzitiva continînd $R^{j}(\theta)$, pentru scopurile noastre poate fi prea mare, deoarece clasele de echivalenta ale lui $R^{j}(\theta)^{T}$ constau din toate acele niveluri de gri care sînt în legatura si pot fi atinse din una în cealalta în relatia $R^{j}(\theta)$. Ne putem astepta ca regularitatea spatiala a multimii imaginilor sa faca toate nivelurile gri atunci când nuantele sînt apropiate pentru θ suficient de mic. Deci, vom propune în continuare alte doua proceduri pentru obtinerea unei relatii de echivalenta mai mica sau mai fina din $R^{j}(\theta)$.

Definim relatia binara $V^{j}(n, m)$ pe Lj prin

$$V^{j}(n, m) = \{(l_{ji}, l_{jk}) \in L_{j} \times L_{j} | n < i \leq m, n < j \leq m\}$$

Vom determina o multime {vn}, n = 0,1,..,Nj(θ) de indici 0 = v0 < v1 < vn<..<vN(θ) = Nj astfel ca relatia de echivalenta F^{j} pe Lj este definita prin :

$$F^{j} = \bigcup_{n=1}^{N(\theta)} V^{j}(v_{n-1}, v_{n})$$

Relatia de echivalenta F^{j} determina partitia H^{j} a lui Lj unde $H^{j} = \{h_{i}^{j}\}, i = 1,2,\ldots, N(\theta)$ si

$$h_{i}^{j} = \{l_{jv_{i-1}+1}, l_{jv_{i-2}+2}, \ldots, l_{jv_{i}}\}, i = 1,2,\ldots, N(\theta)$$

Determinam indicii vn iterativ. Fie v0=0

Presupunem am definit vn-1 si vom defini vn.

Fie a_{k}^{jn} definit prin:

$$a_k^{jn} = \frac{card(R^j(\theta) \cap V^j(v_{n-1}, v_{n-1}+k))}{cardV^j(v_{n-1}, v_{n+1}+k)} \quad k = 1,2,.., N_j - v_{n-1}$$

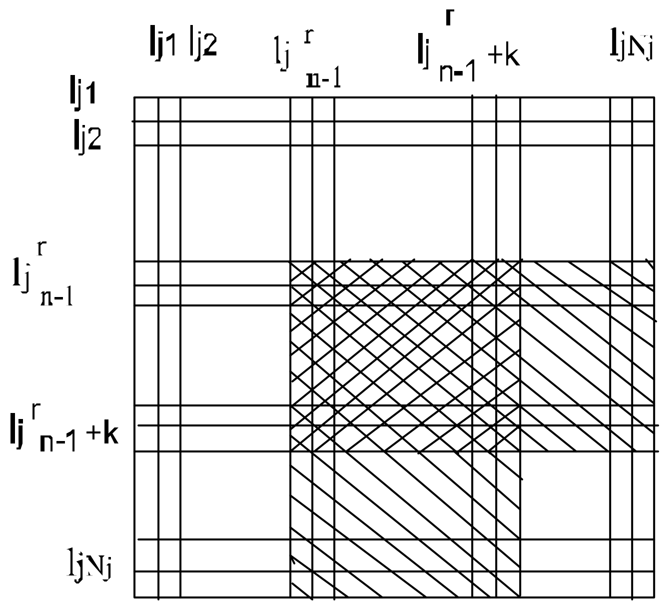
şi d_k^{jn} definit prin:

$$d_k^{jn} = \begin{cases} \dfrac{card(R^j(\theta)^c \cap V^j(v_{n-1}, N_j) \cap V^j(v_{n-1}, v_{n-1}+k)^c \cap V^j(v_{n-1}+k, N_j)^c)}{card(V^j(v_{n-1}, N_j) \cap V^j(v_{n-1}, v_{n-1}+k)^c \cap V^j(v_{n-1}+k, N_j)^c)} \\ \qquad \text{pentru } k = 1,2,.., N_j - V_{n-1} - 1 \\ 1 \qquad \text{pentru } k = N_j - V_{n-1} \end{cases}$$

a_k^{jn} este procentajul de perechi $L_j \times L_j$ care sînt în $R^j(\theta) \cap V^j(v_{n-1}, v_{n-1}+k)$. d_k^{jn} este procentajul din perechi din $L_j \times L_j$ care nu sînt în $R^j(\theta)$ dar sînt în $V^j(v_{n-1}, N_j)$, $V^j(v_{n-1}, v_{n-1}+k)^c, V^j(v_{n-1}+k, N_j)^c \quad (= U)$. Daca multimea $V^j(v_{n-1}, v_{n-1}+k)$ este parte a unei relatii de echivalenta, atunci interpretarea lui a_k^{jn} si d_k^{jn} este ca a_k^{jn} este procentajul de perechi din $R^j(\theta)$, (relativ la $V^j(v_{n-1}, v_{n-1}+k)$ care ar fi în $R^j(\theta)$ si d_k^{jn} este 1 minus procentajul perechilor în $R^j(\theta)$, relativ la U) care nu ar fi în $R^j(\theta)$ si d_k^{jn} este 1 minus procentajul perechilor în $R^j(\theta)$, relativ la U care n-ar fi în $R^j(\theta)$. Deci pentru k= 1,2,..,Nj-vn-1 apropierea relativa sau cât de buna este multimea $V^j(v_{n-1}, v_{n-1}+k)$ poate fi data prin $a_k^{jn} + d_k^{jn}$. Definim vn cel mai mic indice astfel ca:

$$a_{v_n - v_{n-1}}^{jn} + d_{v_n - v_{n-1}}^{jn} \geq a_k^{jn} + d_k^{jn}, k = 1,2,.., N_{j-v_{n-1}}$$

$V^j(v_{n-1}+1, v_n)$ este cea mai buna multime a tuturor multimilor sub conditia de a include relatia de echivalenta F^j.



O procedura alternativa

Dificultatea acestei proceduri este ca daca reordonam multimea Lj astfel ca lji este acum cel mai mare element si ljNj este cel mai mic element, rezultatele obtinute prin folosirea procedurii asupra multimii reordonate nu sînt necesar aceleasi ca rezultatele asupra multimii ordonate initial.

De aceea, vom apela la urmatoarea procedura.

În loc sa obtinem clase de echivalenta secvential în ordinea ordonarii relatiei pe Lj, vom obtine din $R^j(\theta)$, cele mai bune clase de echivalenta prima data si cele mai rele la urma.

Fie $L_j = \{l_{j1}, l_{j2}, .., l_{jNj}\}$ multimea de valori a criteriului Xj si presupunem ca este ordonata prin densitatea gri de la cea mai joasa la cea mai înalta. Fie

$T_0 = \{(n,k) \mid n = 0,1,2,..N_{j-1} {}^o ik = n+1,.., N_j\}$

.Definim coeficientii t_{nk}^j prin

$t_{nk}^j = a_{nk}^j + b_{nk}^j \, pentru (n,k) \in T_0$ unde

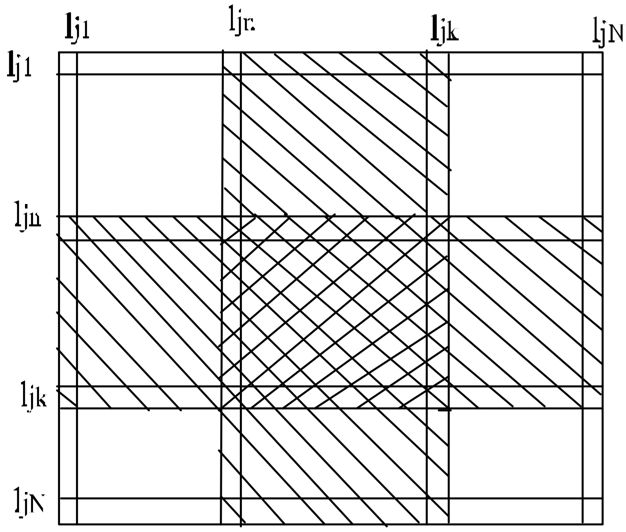
$$a_{nk}^j = \frac{card(R^j(\theta) \cap V^j(n,k))}{card(V^j(n,k))} \quad (n,k) \in T_0$$

$$b_{nk}^j = \begin{cases} \dfrac{card((R^j(\theta)^c \cap (V^j(0,k) \cap V^j(0,n-1)^c \cup (V^j(n-1), N_j) \cap V^j(k, N_j)^c)) \cap V^j(n-1,k)^c)}{card((V^j(0,k) \cap V^j(0,n-1)^c) \cup (V^j(n-1,N_j) \cap V^j(k,N_j)^c) \cap V^j(n-1,k)^{c)}} & \begin{array}{l} \text{pentru } (n,k) \in T_0 \\ n \neq 0 \; {}^o i \; k \neq N_j \end{array} \\ \dfrac{card((R^j(\theta)^c \cap V^j(0,k) \cap V^j(0,n-1)^c \cap V^j(n-1,k)^c)}{card((V^j(0,k) \cap V^j(0,n-1)^c \cap V^j(n-1,k)^c)} & \begin{array}{l} \text{pentru}(n,k) \in T_0 \\ k = N_j, n \neq 0 \end{array} \\ \dfrac{card((R^j(\theta)^c \cap V^j(n,N_j) \cap V^j(k+1,N_j)^c) \cap V^j(n-1,k)^c)}{card((V^j(n,N_j) \cap V^j(k+1,N_j)^c) \cap V^j(n-1,k)^c)} & \begin{array}{l} \text{pentru } (n,k) \in T_0 \\ n = 0 \; {}^o i \; k \neq N_j \end{array} \\ 0 & \text{pentru } (n,k) \in T_0, n = 0, k = N_j \end{cases}$$

a_{nk}^j este fractia tuturor perechilor din $R^j(\theta)$ care ar fi în clase de echivalenta , daca domeniul $V^j(n-1,k)$ este clasa de echivalenta .

b_{nk}^j este fractia tuturor perechilor din $R^j(\theta)$ care n-ar fi în $R^j(\theta)$ daca domeniul $V^j(n-1,k)$ este clasa de echivalenta.

Relatia de echivalenta ceruta F^j este definita prin $F^j = \bigcup_{n=1}^{N(\theta)} W_n$. Submultimile Wn sînt definite iterativ. Fie (s1,l1) oricare pereche de indecsuri satisfacînd $t_{s_1,l_1}^j \geq t_{nk}^j, (n,k) \in T_0$.



Definim atunci:

$W_1 = V^j(s_1, l_1)\ si\ T_1 = \{(n.k) \in T_0 \mid (s_1 \le n \le l_1\}$
$sau\ (s_1 < k \le l_1)\ sau\ (n \le s_1\ si\ k \ge l_1\}$

T1 este multimea tuturor indecsurilor care trebuie respinse din cauza blocului care a fost ridicat. Deci T1 include toate punctele de indici care depasesc cu blocul prezent $V^j(s_1, l_1)$. Daca T1=T0, procedura se opreste si definim F^j=W1 si N(θ) =1.

Presupunem ca am definit Wr-1 si Tr-1. Daca $\bigcup_{n-1}^{r-1} T_n = T_0$ procedura se opreste si definim: $F^j = \bigcup_{n=1}^{r-1} W_n$ si $N(q) = r - 1$. Daca $\bigcup_{n=1}^{r=1} T_n \ne T_0$, fie o pereche de indecsuri $(s_r, l_r) \in T_0 - \bigcup_{n=1}^{r-1} T_n$ care satisface :

$t^j_{s_r l_r} \ge t^j_{nk}, (n,k) \in T_0 - \bigcup_{n=1}^{r-1} T_n$.

Definim $W_r = V^j(s_r, r)$ si $T_r = \{(n.k) \in T_0 \mid s_r \le n \le l_r$ sau $(s_r < k \le l_r)$ sau $(n \le s_r\,^{o} ik \ge l_r)\}$. Daca $\bigcup_{n=1}^{r} T_n = T_0$, procedura este oprita si definim $F^j = \bigcup_{n=1}^{r} W_n\,^{o} iN(q) = r.$

Altfel continuam iterativ. Dupa ce procedura s-a terminat, relatia de echivalenta F^j este definita prin $F^j = \bigcup_{n=1}^{N(\theta)} W_n$. Partitia H^j a lui Lj pe care F^j o determina este definita prin :

$H^j = \{h_i^j\}, i = 1,2,..N(q)$, unde $h_i^j = \{l_{jk} \mid (l_{jk}, l_{jl_i}) \in F^j\}$. h_j^i poate fi echivalent exprimat ca

$h_i^j = Domeniu(W_i) = \{l_{js_i+1}, .., l_{jl_i}\}$.

Eliminarea pragului de cuantificare

Corespunzator manierei pragului spatial de cuantificare, putem de asemenea sa avem o procedura fara prag de cuantificare. Fie matricea: $P^j = (p^j_{ik})$, definita prin:

$$P^j_{ik} = \frac{card\{(z_1, z_2) \in Z_X \times Z_Y | 0 < dist((z_1, z_2), I_J^{-1}(l_{ji})) \le \alpha\,^{o}\ i\ I(z_1, z_2) = l_{jk}\}}{card(Z_x \times Z_y)}$$

Fie factorul de normalizare dr media elementelor de pe diagonala si a primelor r din afara diagonalei matricii P^j.

$$d_r = \frac{\sum_{i=1}^{N_{j-1}} \sum_{k=1}^{r} (p^j_{i,i+k} + p^j_{i+k,i}) + \sum_{i=1}^{N_j} p^j_{ii}}{N_j + 2\sum_{k=1}^{r} (N_j - r)}$$

Fie

$$T_0 = \{(n,k) | n = 0,1,2,.., N_j - 1,^{o}\ i\ k = n+1,.., N_j\}$$

Definim coeficientii t^j_{nk} prin:

$$t^j_{nk} = a^j_{nk} + b^j_{nk} \text{ pentru } (n,k) \in T_0, \text{ unde } a^j_{nk} = \sum_{u=n}^{k}\sum_{v=n}^{k} p^j_{uv} \text{ iar}$$

$$b^j_{nk} = 2[\sum_{n=1}^{N_j-1}\sum_{v=n}^{k}(d_k - p^j_{uv}) + \sum_{u=k+1}^{N_j}\sum_{v=n}^{k}(d_r - p^j_{uv})]$$

Relatia de echivalenta ceruta F^j este definita prin $F^j = \bigcup_{n=1}^{N(d_r)} W_n$. Submultimile Wr sînt definite în aceeasi maniera ca si mai înainte.
Daca sînt obtinute mai multe clase de echivalenta, **r** poate fi selectat mai mare. Daca sînt obtinute mai putine clase de echivalenta, atunci **r** poate fi ales mai mic.

Bibliografie

1.Emil Stan "Modelarea dinamica a sistemelor", Editura RDA, 1998
2.Emil Stan "Robust automatic classification", Information Society, INFOREC Printing House, 2001
3.Emil Stan "Clasificarea prin proiectii", Informatica Economica, nr.1, 2001
4. W. D. Ficher "On Grouping for Maximum Homogenity", vol 53, Dec. 1998, Journal of America St. Ass.