

Clasificarea prin proiectie

Conf.dr. Emil STAN

Facultatea de Drept, Academia de Politie "Alexandru Ioan Cuza", Bucuresti

Unul din instrumentele de mare importanta ale economistilor, dar nu numai al acestora ramâne clasificarea automata. În lucrarea de fata se propune un algoritm care foloseste tehnicile calcului diferential si ale statisticii matematice pentru obtinerea claselor partitiei.

Cuvinte cheie: clasificare automata, distributie normala, partitie.

Procedurile obisnuite de clasificare consuma mult timp calculator. Frontierele dintre clustere sunt neliniare. **Clasificarea prin proiectii** propusa în aceasta lucrare este o metoda rapida, însa frontierele dintre clustere sunt liniare pe portiuni. Ideea de baza este de a proiecta masuratorile pe un subspatiu mono-dimensional si apoi a determina cel mai bun punct în care sa despartim clusterele. Proiectia inversa a acestui punct, defineste hiperplanul de separare.

Când datele proiectate prezinta o **distributie multimodala**, atunci exista cel mai bun punct de separare si este localizat în cea mai adâncă "vale" dintre perechile de moduri. Când datele proiectate prezinta o distributie unimodala, nu exista punct de separare.

Multimea vectorilor poate fi împartita în doua submultimi. O submultime va cuprinde punctele mai mari decât punctul de separare si cealalta va fi complementara.

Dupa o prima separare, fiecare submultime se poate trata analog. Acestea sunt, de asemenea, proiectate pe subspatiul mono-dimensional optim pentru separare.

Aceasta procedura se continua pâna când fiecare submultime masura are o **distributie unimodala** si deci nu se mai poate separa.

Vom proiecta multimea pe subspatiul uni-dimensional pe care submultimea masura are cea mai mare împrastiere. Pentru a determina punctul de separare ne vom folosi de deviatia absoluta medie normalizata.

Pentru fiecare punct de separare posibil,

examinam deviatia absoluta medie normalizata a tuturor vectorilor masura mai mici ca el.

Cel mai bun punct de separare este fixat în "valea" a doua moduri. Acesta este locul în care derivata a II-a a deviatiei absolute este cea mai mare.

Pentru expunerea ideilor notam $S^0 = (g_1, g_2, \dots, g_m)$ vectorii masura. Fiecare g_i este un element al lui R^N . S^{00} este acea parte a secventei de date S^0 cu imaginea proiectata mai mica decât primul punct de separare po. Vom spune ca S^{00} este o parte a hiperplanului H_0 care este definit de proiectia inversa a lui po. S^{01} este acea parte a lui S^0 cu proiectia mai mare decât po. Deci toate masuratorile din S^{01} stau pe partea 1 a hiperplanului H_0 . Analog, pentru orice secventa $(n_1, n_2, \dots, n_k)$ de cifre binare, definim $S^{n_1 n_2 \dots n_k}$ ca elementele din S^0 care stau pe partea n_k a hiperplanului $H_{n_1, n_2, \dots, n_{k-1}}$, partea n_{k-1} a hiperplanului $H_{n_1, n_2, \dots, n_{k-2}}$, si partea n_1 a hiperplanului H_0 .

Definim iterativ:

S^0 secventa originala. Pentru orice secventa $(n_1, n_2, \dots, n_k)$ de cifre binare consideram separarea secventei $S^{n_1 n_2 \dots n_k} = (\gamma_1, \gamma_2, \dots, \gamma_t)$. Fie $X^{n_1 n_2 \dots n_k}$ matricea de date pentru $S^{n_1 n_2 \dots n_k}$

Deci,

$$X^{n_1, n_2, \dots, n_k} = \begin{pmatrix} \gamma_{11} & \gamma_{12} & \dots & \gamma_{1N} \\ \gamma_{21} & \gamma_{22} & \dots & \gamma_{2N} \\ \dots & \dots & \dots & \dots \\ \gamma_{t1} & \gamma_{t2} & \dots & \gamma_{tN} \end{pmatrix}$$

unde $\gamma_i = \begin{pmatrix} \gamma_{i1} \\ \gamma_{i2} \\ \dots \\ \gamma_{iN} \end{pmatrix}$ este a i -a masura în

$S^{n_1, n_2, \dots, n_k}$

Spatiul optimal uni-dimensional este acel subspatiu cu împrastierea cea mai mare.

Ca mai înainte, definim matricea împrastierii :

$$D^{n_1, n_2, \dots, n_k} = (X^{n_1, n_2, \dots, n_k} - 1_t \cdot \bar{g}') \bullet (X^{n_1, n_2, \dots, n_k} - 1_t \cdot \bar{g})$$

unde

$$\bar{g} = \frac{(X^{n_1, n_2, \dots, n_k})' * 1_t}{t}$$

Dacă $r^{n_1, n_2, \dots, n_k}$ este vectorul $N \times 1$ al direcției cosinus pentru subspatiul optimal uni-dimensional, atunci împrastierea proiectată $d^{n_1, n_2, \dots, n_k}$ este:

$$d^{n_1, n_2, \dots, n_k} = (r^{n_1, n_2, \dots, n_k})' * D^{n_1, n_2, \dots, n_k} * r^{n_1, n_2, \dots, n_k}$$

Pentru a determina $r^{n_1, n_2, \dots, n_k}$ maximizăm $d^{n_1, n_2, \dots, n_k}$, atunci când $(r^{n_1, n_2, \dots, n_k})' * r^{n_1, n_2, \dots, n_k} = 1$, folosind multiplicatorii Lagrange.

Definim lagrangianul:

$$\Phi = (r^{n_1, n_2, \dots, n_k})' * D^{n_1, n_2, \dots, n_k} * r^{n_1, n_2, \dots, n_k} - \lambda ((r^{n_1, n_2, \dots, n_k})' * r^{n_1, n_2, \dots, n_k} - 1)$$

Ținând seama de simetria lui $D^{n_1, n_2, \dots, n_k}$ obținem :

$$\frac{\partial \Phi}{\partial r^{n_1, n_2, \dots, n_k}} = D^{n_1, n_2, \dots, n_k} * r^{n_1, n_2, \dots, n_k} - \lambda * r^{n_1, n_2, \dots, n_k}$$

Punând condiția ca aceste derivate parțiale să fie 0, obținem:

$$D^{n_1, n_2, \dots, n_k} * r^{n_1, n_2, \dots, n_k} = \lambda r^{n_1, n_2, \dots, n_k}$$

Deci $r^{n_1, n_2, \dots, n_k}$ va fi un vector propriu al lui $D^{n_1, n_2, \dots, n_k}$.

Atunci împrastierea proiectată va fi:

$$d^{n_1, n_2, \dots, n_k} = (r^{n_1, n_2, \dots, n_k})' * (D^{n_1, n_2, \dots, n_k} * r^{n_1, n_2, \dots, n_k}) =$$

$$(r^{n_1, n_2, \dots, n_k})' * (\lambda * r^{n_1, n_2, \dots, n_k}) = \lambda$$

$r^{n_1, n_2, \dots, n_k}$ trebuie să fie vectorul propriu al lui $D^{n_1, n_2, \dots, n_k}$ cu cea mai mare valoare proprie.

Fie $\omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \dots \\ \omega_t \end{pmatrix} = X^{n_1, n_2, \dots, n_k} * r^{n_1, n_2, \dots, n_k}$ vectorul de date proiectat.

Fie $\omega_{\min} = \min_{1 \leq i \leq t} \omega_i$; $\omega_{\max} = \max_{1 \leq i \leq t} \omega_i$

$$\omega_{\min} = \min_{1 \leq i \leq t} \omega_i; \omega_{\max} = \max_{1 \leq i \leq t} \omega_i$$

Vectorul datelor proiectat va fi y dat de relația :

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_t \end{pmatrix} \text{ unde } y_i = \frac{\omega_i - \omega_{\min}}{\omega_{\max} - \omega_{\min}} (t - 1) + 1,$$

Fie z vectorul y cu componentele ordonate

$$z = \begin{pmatrix} z_1 \\ z_2 \\ \dots \\ z_t \end{pmatrix} \text{ Devierea absoluta medie pentru primele } n \text{ puncte o definim prin:}$$

$$dn(z) = \frac{1}{n} \sum_{i=1}^n |z_i - \bar{z}_n| \text{ unde } \bar{z}_n = \frac{1}{n} \sum_{i=1}^n z_i$$

Un sir distribuit perfect uniform ordonat

$$\text{crescator ia forma : } z = \begin{pmatrix} 1 \\ 2 \\ \dots \\ m \end{pmatrix} \text{ astfel ca}$$

$$\bar{z}_m = \frac{1}{m} \sum_{n=1}^m n = \frac{1}{m} * \frac{m(m+1)}{2} = \frac{m+1}{2}$$

Să presupunem ca m este par. Atunci :

$$\begin{aligned}
d(z) &= \frac{1}{m} * \sum_{i=1}^m |z_i - \bar{z}| = \frac{1}{m} * \sum_{i=1}^m |i - \frac{m+1}{2}| \\
&= \frac{1}{m} * \sum_{i=1}^{\frac{m}{2}} (\frac{m+1}{2} - i) + \frac{1}{m} * \sum_{i=\frac{m}{2}+1}^m (i - \frac{m+1}{2}) \\
&= \frac{2}{m} * \sum_{i=1}^{\frac{m}{2}} (\frac{m+1}{2} - i) = \frac{2}{m} * \{ \frac{m+1}{2} * \frac{m}{2} - \frac{m}{2} * (\frac{m}{2} + 1) \} = \frac{m}{4}
\end{aligned}$$

Dacă m este impar:

$$\begin{aligned}
d(z) &= \frac{1}{m} * \sum_{i=1}^m |z_i - \bar{z}| = \frac{1}{m} * \sum_{i=1}^m |i - \frac{m+1}{2}| \\
&= \frac{1}{m} * \sum_{i=1}^{\frac{m-1}{2}} (\frac{m+1}{2} - i) + \frac{1}{m} * |\frac{m+1}{2} - \frac{m+1}{2}| + \\
&+ \frac{1}{m} * \sum_{i=\frac{m+3}{2}}^m (i - \frac{m+1}{2}) = \frac{2}{m} * \sum_{i=1}^{\frac{m-1}{2}} (\frac{m+1}{2} - i) \\
&= \frac{m^2 - 1}{4 * m}
\end{aligned}$$

În continuare, fie $u = \begin{pmatrix} 1 \\ 2 \\ \vdots \\ n \end{pmatrix}$. Am arăta deci:

$$dn(u) = \begin{cases} \frac{n}{4} & \text{dacă } n \text{ este par} \\ \frac{n^2 - 1}{4 * n} & \text{dacă } n \text{ este impar} \end{cases}$$

Normalizăm prin deviațiile unui astfel de distribuții uniforme din cauza că "vârte" pe care dorim să le găsim sunt "vârte" relative la o distribuție uniformă. Atunci, definim deviațiile absolute în medie normalizate prin:

$$\delta_n(z) = d_n(z) - d_n(u), \quad n = 1, 2, \dots, t$$

Pentru separare, găsim cel mai mic index j care maximizează diferența a II-a. Atunci j satisface:

$$(d_j(z) - d_{j-1}(z)) - (d_{j-1}(z) - d_{j-2}(z)) \geq (d_n(z) - d_{n-1}(z)) - (d_{n-1}(z) - d_{n-2}(z)), n = 1, 2, 3, \dots, t$$

Fie ϑ un prag de separare dat. Atunci dacă:

$$d_j(z) - 2d_{j-1}(z) + d_{j-2}(z) \geq \vartheta, \quad \text{efectuăm}$$

separarea prin punctul $p_{n1, n2, \dots, nk} = \frac{z_{j-2} + z_{j-1}}{2}$

Atunci $S^{n1, n2, \dots, nk0}$ este cel mai mare subsir al lui $S^{n1, n2, \dots, nk}$ a cărui imagine este mai mică decât $p_{n1, n2, \dots, nk}$ și $S^{n1, n2, \dots, nk1}$ este cel mai mare subsir al lui $S^{n1, n2, \dots, nk}$ a cărui imagine proiectată este mai mare decât $p_{n1, n2, \dots, nk}$. Dacă indexul j nu satisface cerința anterioară relativă la depășirea pragului ϑ atunci $S^{n1, n2, \dots, nk}$ nu este separat mai departe și clusterul său este definit prin toate măsurile spațiului N dimensional euclidian care stau în a n_k parte a hiperplanului $H_{n1, n2, \dots, nk-1}$ parte a lui $H_{n1, n2, \dots, nk-2, \dots, n1}$ parte a lui H_0 .

Întrucât procedura începe cu întreg spațiul euclidian, divizându-l în două prin hiperplanul H_0 și apoi continuă prin subdiviziunea fiecărei părți, trebuie ca reuniunea tuturor subdiviziunilor obținute să fie întreg spațiul euclidian. Clusterelor sunt disjuncte și acoperă spațiul de măsură și, prin urmare, sunt celule ale partiției spațiului de măsură.

Metoda prezentată diferă de metodele clasice de clasificare, cum sunt metoda "cluster analysis" sau metoda centrelor de greutate, utilizând tehnicile calculului diferențial și ale statisticii matematice. Clasificarea obținută va fi optimă în raport cu criteriul ales.

Bibliografie

[1] Emil STAN, Modelarea dinamică a sistemelor, Editura RDA, București, 1998