Analiza de cluster

Algoritmii de clusterizare sau clasificare nesupervizată sunt utilizați pentru a determina grupările naturale ale datelor sau pentru a furniza o împărțire convenabilă a datelor în grupuri.

Spre deosebire de clasificarea supervizată (analiza discriminantă), în clasificarea nesupervizată nu există o informație prealabilă despre grupele în care se va face clasificarea. Analiza de cluster este o metodă exploratorie. Numărul grupelor este subordonat scopului analizei.

Datele supuse analizei sunt valori ale relațiilor dintre indivizii și variabilele aflate în studiu, luate câte două - distanțe sau disimilarități. Înregistrarea acestor valori ia forma matricelor de distanță. Așadar, valorile supuse analizei reprezintă deja rezultatul unor calcule care au ca rezultat măsuri de disimilaritate dintre obiecte (instanțe sau variabile).

Grupele formate prin clasificare conțin indivizi asemănători intre ei, cu disimilarități mici. Privită din această perspectivă, a formării de grupe omogene, analiza de cluster este o metodă de sinteză informațională, așa cum este și analiza în componente principale, doar că se referă în principal la instanțe și nu la variabile. O grupă omogenă poate fi reprezentată în diverse analize printr-un singur individ: centrul grupei.

Există mai multe motive pentru care o grupare a datelor este necesară:

- *Identificarea trăsăturilor fundamentale ale datelor*. Se studiază relațiile semnificative existente între date.
- *Obținerea unor reprezentări avantajoase în efectuarea analizelor*
- *Stocare și regăsire rapidă a informației*. În acest sens este preocupantă îmbunătățirea vitezei de acces prin furnizarea unor strategii de rutare pentru informația stocată. Eficiența grupării se măsoară prin eficiența timpului de regăsire a unei părți dorite din ansamblul datelor.

Descrierea algoritmilor de clasificare

Algoritmii de clusterizare se împart în următoarele grupe:

- algoritmi ierarhici;
- algoritmi de partiționare (KMeans)
- algoritmi aglomerativi (K-nn)
- algoritmi de tip grid (OCluster)

Algoritmi ierarhici

Fie $\Omega = \{w_1, w_2, ..., w_n\}$ mulțimea obiectelor (indivizi/instanțe sau variabile) aflate în analiză. O ***ierarhie***, notată cu *H*, este un ansamblu ordonat de mulțimi, formate din elemente ale mulțimii Ω, agregate la un anumit nivel, si care are proprietățile:

1. $\Omega \in H$, adică submulțimea agregată la nivelul cel mai de sus conține toți indivizii;
2. pentru orice w_i, i=1,n (n este numărul de obiecte), există $\{w_i\} \in H$ care formează submulțimile de bază, terminale;
3. pentru orice $h, h' \in H$ există implicația: $h \cap h' \neq \emptyset \Rightarrow h \subset h'$ sau $h' \subset h$.

Outputul grafic al algoritmilor ierarhici este ***graficul dendrogramă***. Una dintre axele graficului este axa distanțelor. Cealaltă axă este axa obiectelor. Graficul evidențiază distanțele de agregare din ierarhie.

Exemplu. În figura următoare este construită o ierarhie după distanța euclidiană pentru 5 puncte din plan și graficul aferent

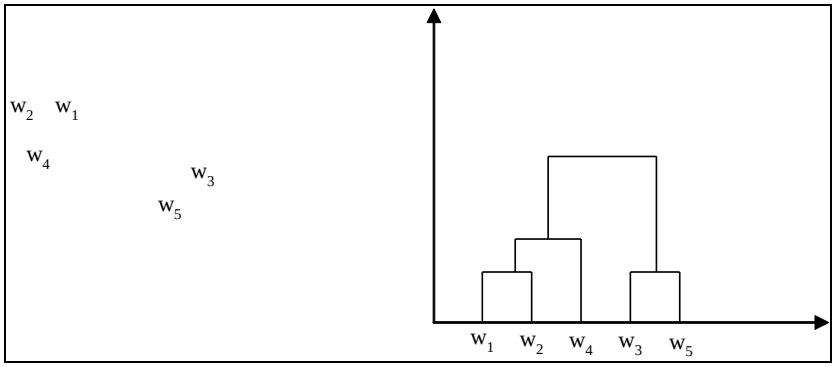


Figura 1. Exemplu de ierarhie

Iată pașii prin care se construiește o ierarhie:

***Intrări*:** un set de $n(n\text{-}1)/2$ distanțe și un set de n mulțimi $h_i = \{w_i\}$, $i = 1,n$, aflate la nivelul cel mai de jos de agregare

***Pasul* 1**: se selectează mulțimile care vor fi agregate: h_i și h_k

***Pasul* 2:** se adună obiectele mulțimilor h_i și h_k și se înlocuiesc cu un nouă mulțime $h' = h_i \cup h_k$ actualizându-se și distanțele între noua mulțime h' și celelalte mulțimi

***Pasul 3*:** Atâta timp cât rămân cel puțin două mulțimi, se revine la pasul 1.

O ierarhie se construiește în n-1 etape/iterații. La fiecare iterație are loc o joncțiune, două mulțimi dispar în locul lor apărând una nouă formată prin reuniune. Se poate spune că după fiecare joncțiune se obține o nouă repartizare a obiectelor în mulțimi. O astfel de repartizare este numită ***partiție***.

Prima partiție corespunde nivelului cel mai de jos de repartizare: $P_1 = \{\{w_1\},\{w_2\}, ..., \{w_n\}\}$. Următoarele partiții vor avea n-1, n-2, ..., 1 elemente. Ultima partiție este $P_n = \{\{w_1,w_2, ...,w_n\}\}$.

Partițiile pot fi evidențiate in graficul dendrogramă prin secționări paralele cu axa distanțelor. Numărul de mulțimi dintr-o partiție este egal cu numărul de brațe intersectate de secțiune. In imaginile de mai jos sunt evidențiate diverse partiții obținute prin secționare.

Un ***cluster*** este definit printr-o mulțime de obiecte din partiție sau din ierarhie.

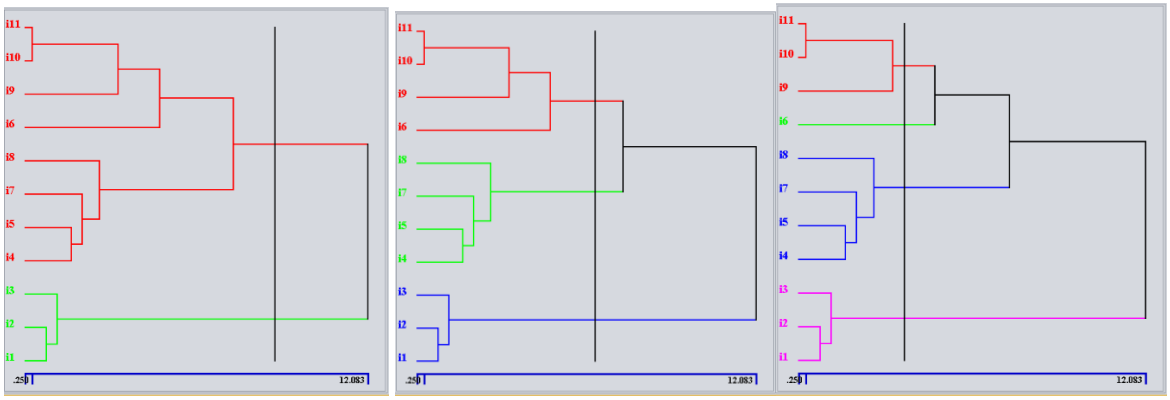


Figura 2. Partiții cu 2,3 și 4 clustere

Un algoritm general de clusterizare ierarhică poate fi scris în pseudocod astfel:

```
Procedure Ierarhie(D;H)
array D(n,n),P(n)
List[] H
for i=1,n
        P[i] = {i}
        call add(H,{i})
endfor
for i=1,n-1
        call Select(D,P;k,j)
        P[k] = P[k] + P[j]
        P[j] = {}
        call add(H,P[k])
        call Actualizare(D,k,j)
        call Afisare(P)
endfor
return
end
```

Unde:

D - este matricea de distanțe

P - este un vector de liste de mulțimi (clustere), utilizat pentru memorarea unei partiții. În listă sunt memorate valorile index ale obiectelor/instanțelor din mulțime (indexul in tabelul de observații).

H - este o listă de mulțimi (clustere) și reprezintă ierarhia. În final *H* va avea 2*n*-1 elemente (*n* clustere de start, *n*-1 clustere obținute în cele *n*-1 joncțiuni ale algoritmului).

Procedurile *Select* și *Actualizare* individualizează fiecare implementare. Procedura *Select* implementează principiul de grupare iar procedura *Actualizare* recalculează distanțele după joncțiune. Procedura *Add* adaugă un cluster la ierarhie.

Toate metodele de grupare ierarhică urmează acest algoritm. Diferențele sunt date de modul în care se aleg clusterele care joncționează la un moment dat și de modul în care sunt actualizate distanțele după joncționare.

Metode de grupare ierarhică

1. ***Gruparea prin legătură simplă***. *Selecția* clusterelor se face prin distanța minimă. *Actualizarea* distanțelor după joncționare se face după relația:

$$d^2(h_s \cup h_r, h_i) = Minim\{d^2(h_s,h_i), d^2(h_s,h_i)\},$$

unde h_s şi h_r sunt clusterele selectate pentru joncţionare, iar h_i este unul dintre clusterele rămase.

Exemplu.

Sunt prezentate în continuare agregările şi construirea dendogramei pornind de la un tabel de distanţe pentru 5 indivizi.

	1	2	3	4	5
1	0	4	9	5	8
2	4	0	6	3	6
3	9	6	0	6	3
4	5	3	6	0	5
5	8	6	3	5	0

Distanţele iniţiale

	1	2∪4	3	5
1	0	4	9	8
2∪4	4	0	6	5
3	9	6	0	3
5	8	5	3	0

Distanţele dintre 2-4 şi restul

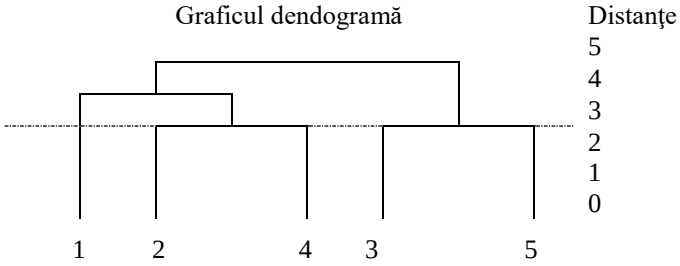
	1	2∪4	3∪5
1	0	4	8
2∪4	4	0	5
3∪5	8	5	0

Distanţele dintre 3-5 şi restul

	1∪2∪4	3∪5
1∪2∪4	0	5
3∪5	5	0

Distanţele dintre 1-2-4 şi 3-5

Graficul dendogramă



Dendrograma

2. ***Gruparea prin legătură completă***. *Selecţia* clusterelor se face prin distanţa minimă. *Actualizarea* distanţelor după joncţionare se face după relaţia: $d^2(h_s \cup h_r, h_i) = \text{Maxim}\left\{d^2(h_s, h_i), d^2(h_s, h_i)\right\}$.

3. ***Gruparea prin legătură medie***. *Selecţia* clusterelor se face prin distanţa minimă. *Actualizarea* distanţelor după joncţionare se face după relaţia:

$$d^2(h_s \cup h_r, h_i) = \begin{cases} \dfrac{d^2(h_s, h_i) + d^2(h_s, h_i)}{2} & \text{Medie neponderată} \\ \dfrac{n_s \cdot d^2(h_s, h_i) + n_r \cdot d^2(h_r, h_i)}{n_s + n_r} & \text{Medie ponderată} \end{cases}$$

unde:

n_r - numărul de indivizi din h_r

n_s - numărul de indivizi din h_s.

4. ***Metoda centroidă***. *Selecţia* clusterelor se face prin distanţa minimă. *Actualizarea* distanţelor după joncţionare se face după relaţia:

$$d^2(h_s \cup h_r, h_i) = \begin{cases} \dfrac{d^2(g_s, g_i) + d^2(g_r, h_i)}{2} & \text{Medie neponderată} \\ \dfrac{n_s \cdot d^2(g_s, g_i) + n_r \cdot d^2(g_r, g_i)}{n_s + n_r} & \text{Medie ponderată} \end{cases}$$

unde g_s şi g_r sunt centrii clusterelor h_s şi h_r.

5. ***Metoda varianţei minime sau Ward***. Se bazează pe descompunerea varianţei totale în varianţă intra-clase şi varianţă inter-clase.

Fie q numărul grupelor (clusterelor) existente la un moment dat. Varianţele totală, intra-clase şi inter-clase sunt definite astfel:

$T = \dfrac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_j\right)^2$ - varianţa totală,

$W = \sum_{k=1}^{q}\dfrac{n_k}{n} \cdot \dfrac{1}{n_k}\sum_{i \in h_k}\sum_{j=1}^{m}\left(x_{ij} - g_{kj}\right)^2$ - varianţa intra-clase,

B = $\sum_{k=1}^{q} \frac{n_k}{n} \sum_{j=1}^{m} \left(g_{kj} - \bar{x}_j\right)^2$ - varianța inter-clase,

unde X este tabelul de observații cu termenul general x_{ij}, $\overline{x_j}$ este media generală pentru variabila X_j, g_{kj} este media variabilei X_j pentru grupa k.

Varianța totală este constantă indiferent cum se împart instanțele pe clustere. Variațele intra-clasă și inter-clasă sunt însă dependente de modul în care se împart instanțele pe clustere. O împărțire este cu atât mai bună cu cât clusterele sunt mai omogene, adică varianța intra-clasa va fi mai mică iar varianța inter-clasă va fi mai mare.

La nivelul cel mai de jos, când fiecare din cele n instanțe constituie câte un cluster, varianța intra-clasă este nulă în timp ce varianța inter-clasă este maximă. Dacă se agregă două clustere, varianța intra-clasă crește (clusterul format fiind mai puțin omogen) iar varianța inter-clasă scade. Când toate instanțele sunt grupate într-un singur cluster, varianța intra-clasă crește la valoarea maximă în timp ce varianța inter-clasă scade la 0.

Presupunem că la un moment dat sunt agregate două clustere h_r și h_s. Noul cluster conține n_r+n_s instanțe.

Varianța intra-clasa crește cu valoarea Δ_{rs} obținută prin agregarea clusterelor h_r și h_s: $\Delta_{rs} = \frac{n_r n_s}{n_r + n_s} \cdot d^2(g_s, g_r)$.

După criteriul Ward se agregă la o etapă oarecare acele două clustere h_r și h_s astfel încât Δ_{rs} să fie minimă.

Alegerea numărului optim de clustere

Criterii de alegere a numărului de clustere:

- *Natura problemei*. În funcție de scopul urmărit în analiză, numărul de grupe poate fi mai mare sau mai mic. Decizia aparține specialistului din domeniul în care este efectuată analiza.
- *Variația distanței de agregare*. Se alege partiția care corespunde diferenței maxime de distanță.

În figura de mai jos este aplicat criteriul variației distanței de agregare.

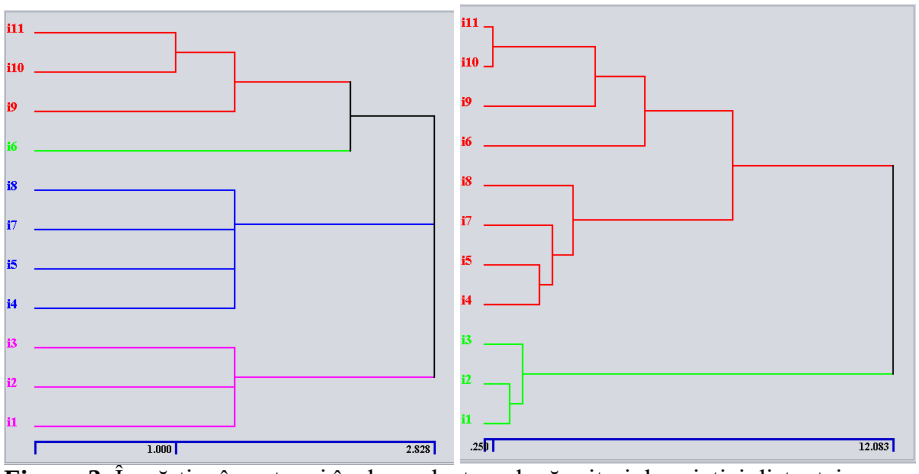


Figura 3. Împărțire în patru și în doua clustere după criteriul variației distanței