# Aggregation Algorithms in Heterogeneous Tables

Conf.dr. Felix FURTUNĂ, prof.dr. Ion IVAN, conf.dr. Marian DÂRDALĂ
Catedra de Informatică Economică, ASE Bucureşti

*The heterogeneous tables are most used in the problem of aggregation. A solution for this problem is to standardize these tables of figures. In this paper, we proposed some methods of aggregation based on the hierarchical algorithms.*
***Keywords**: Heterogeneous table, aggregation, algorithms, tree clustering, joining.*

## 1 Serii şi tabele de date

O ***serie de date*** $X$ având $n$ termeni se descrie ca un şir de forma $x_1,x_2,...,x_n$. În cazul în care toţi termenii aparţin unui domeniu $D$, seria $X$ este o serie omogenă. De exemplu, seria de termeni constituită pentru a înregistra înălţimea persoanelor dintr-o colectivitate $P$ descrisă în Tabelul 1 este o ***serie omogenă***.

**Tabelul 1**

| Persoana | Înălţimea |
|---|---|
| $P_1$ | 176 |
| $P_2$ | 168 |
| $P_3$ | 184 |
| ... | |
| $P_n$ | 165 |

Pentru valorile din Tabelul 1 există un singur domeniu de definire a colectivităţii $P$, format din elementele $P_1,P_2, ..., P_n$ , şi un singur codomeniu $D_1$ al funcţiei $f$ care asociază fiecărui individ înălţimea sa:

$$f:P\rightarrow D_1 = [A,B]\cap N,$$

unde $A$ = min $\{x_1,x_2,...,x_n\}$ şi $B$ = min $\{x_1,x_2, ..., x_n\}$.

Aceeaşi serie poate avea pentru unii indivizi valori ale caracteristicii *Înălţimea* de tipul *scund*, *mediu*, *înalt*, după cum se poate observa în Tabelul 2.

**Tabelul 2**

| Persoana | Înălţimea |
|---|---|
| $P_1$ | 176 |
| $P_2$ | mediu |
| $P_3$ | 184 |
| $P_4$ | scund |
| ... | |
| $P_n$ | 165 |

În această situaţie, pe lângă domeniul $D_1$ se defineşte domeniul $D_2$ = {*scund*, *mediu*, *înalt*}, iar unele elemente ale colectivităţii $P$ formând subcolectivitatea $P'$ sunt descrise cu valori din $D_1$, iar alte elemente, formând subcolectivitatea $P''$ sunt descrise cu valori din domeniul $D_2$. Pentru domeniul $D_2$ avem funcţia $h$ care asociază fiecărui individ din subcolectivitatea $P''$, o valoare din $D_2$:

$$h:P''\rightarrow D_2.$$

O astfel de serie se numeşte ***serie neomogenă***.

Într-un ***tabel de date*** sunt înregistrate mai multe caracteristici ale aceleiaşi populaţii. Notăm cu $C_1$, $C_2$, ..., $C_m$ cele $m$ caracteristici măsurate la $n$ indivizi aparţinând populaţiei. Tabelele de date sunt de forma $n\times m$. Pentru fiecare caracteristică $C_j$ se defineşte un set de domenii $D_1^j, D_2^j,..., D_{m_j}^j$ disjuncte, rezultând modalităţi foarte diferite de completare şi de definire a elementelor unei mulţimi. În această situaţie avem un tabel de date neomogen.

Tabelele de date sunt complete dacă la intersecţia liniei $i$ şi a coloanei $j$ se află un element aparţinând unui domeniu pentru care a fost definit tabelul respectiv. O serie de date este completă dacă provine dintr-un tabel de date complet.

Sunt situaţii în care nu s-au efectuat măsurători pentru nivelurile unor caracteristici, caz în care în tabel fie se lasă liber, fie se trece un simbol cu semnificaţia de vid (de exemplu o linie " - ").

Dacă în domeniul $D_j$ se adaugă elementul vid, se obţine domeniul extins $D_j^e$ . Seriile de date definite pe domeniul $D_j^e$ sunt cu un grad de generalitate mai mare, întrucât includ şi seriile de date incomplete.

Prin concatenarea de serii de date incomplete la un tabel de date complet, referitoare la aceeaşi populaţie, se obţine un tabel de date incomplet. Transformarea tabelului de date incomplet în tabel de date complet impune reluarea în condiţii identice a procedurilor de măsurare pentru a completa datele absente.
Un exemplu de tabel de date neomogen complet definit este următorul:

**Tabelul 3**

| Cod | *Marca* | *Preţ* | *Consum* | *Garanţie* | *CapacitateCilindrică* |
|---|---|---|---|---|---|
| C1 | Renault Clio | 8500 | 7 | 3 | 1400 |
| C2 | Dacia 1300 | 150000000 | 8.5 | 18 | 1400 |
| C3 | Citroen C3 | 9000 | moderat | 3 | medie |
| C4 | Audi A80 | 3500 | mic | 100000 | mare |
| ... | | | | | |
| $C_n$ | WW Golf | 200000000 | mare | 150000 | 2000 |

În tabelul 3 sunt înregistrate date despre maşinile de vânzare ale unei societăţi de profil. Caracteristicile maşinilor cuprind elemente din mai multe domenii. De exemplu, pentru caracteristica *Garanţie* avem domeniul $D_1 = \{0,1,2,3\}$ reprezentând numărul anilor de garanţie, domeniul $D_2 = [0,36] \cap N$ reprezentând numărul lunilor de garanţie, domeniul D3 = $[0,500000] \cap N$ care specifică numărul kilometrilor parcurşi pentru care există garanţie.

## 2. Omogenizarea tabelelor de date

O primă modalitate de omogenizare propune utilizarea unor coeficienţi de omogenizare, $a_1,a_2,...,a_n$ de forma:

$$J_i^{(1)} = \sum_{j=1}^{n} a_j x_{ij}\ ,\ J_i^{(2)} = \sum_{j=1}^{m} \frac{a_j}{x_{ij}}\ ,\ J_i^{(3)} = \sum_{j=1}^{m} x_{ij}^{a_j}\ ,$$

$$J_i^{(4)} = \sqrt[A]{\prod_{j=1}^{m} x_{ij}^{a_j}}\ ,$$

unde $x_{ij}$ reprezintă valoarea înregistrată la individul $i$ pentru caracteristica $C_j$, iar $A = \sum_{j=1}^{m} a_j$ .

O altă modalitate de omogenizare a datelor din tabelele neomogene o constituie realizarea unei proiecţii în spaţiul mărimilor adimensionale, iar normalizarea reprezintă o astfel de proiecţie. Noul tabel de date omogene conţine elementele $x_{ij}^{'}$ obţinute prin relaţia:

$$x_{ij}^{'} = \frac{x_{ij} - x\,min_j}{x\,max_j - x\,min_j},$$

unde $x\,min_j = \min\{x_{1j},x_{2j},...,x_{nj}\}$ iar $xmax_j = \max\{x_{1j},x_{2j},...,x_{nj}\}$.

În cazul în care domeniile pe care sunt definite valorile caracteristicilor nu sunt incluse in R, este necesară punerea în corespondenţă a valorilor cu valori din mulţimea numerelor reale.

Se consideră un domeniu $D_k^j = \{d_1^j, d_2^j, ..., d_{k_j}^j\}$ al caracteristicii $C_j$, unde $k_j$ reprezintă numărul de cuvinte care defineşte domeniul $D_k^j$ . Domeniul $D_k^j$ este de fapt un vocabular, iar $d_1^j, d_2^j, ..., d_{k_j}^j$ sunt cuvintele din respectivul vocabular.

Dacă se defineşte o procedură care să pună în corespondenţă cuvintele vocabularului $D_k^j$ cu un domeniu $D_k^{j*}$ unde $D_k^{j*} = \{v_1^j, v_2^j, ..., v_{k_j}^j\}$, cu $v_i^j \in R$ , $i$=1,2,...,$k_j$, atunci înseamnă că s-a găsit modalitatea de a omogeniza tabelul de date. În cazul în care se definesc proceduri pentru a soluţiona problema pentru toate vocabularele cu care s-a construit tabloul, spunem că s-a obţinut un tabel de date numerice a cărui omogenizare se realizează printr-o proiecţie oarecare.

Punerea în corespondenţă a cuvintelor unui vocabular $W = \{w_1, w_2, ..., w_h\}$ în care $h$ este numărul cuvintelor, se realizează astfel:

- se consideră o mulţime de specialişti $S_1,S_2,...,S_g$, unanim recunoscuţi în domeniul considerat;
- se acordă note de la 1 la $h$ cuvintelor vocabularului W, construindu-se un tablou cu g

linii şi h coloane;

- se calculează ponderile $p_j = \frac{\sum_{i=1}^{g} n_{ij}}{\sum_{i=1}^{g}\sum_{j=1}^{h} n_{ij}}$,

$j$=1,2,...,$h$, asociate cuvintelor vocabularului W;

- se înlocuieşte în tabel cuvântul $w_j$ cu ponderea lui $p_j$.

O a treia metodă de omogenizare este generarea de numere pseudoaleatoare. Dacă într-un tabel de date neomogene apar calificative (de exemplu: foarte bine, bine, satisfăcător, nesatisfăcător), se va produce omogenizarea prin generare de numere pseudoaleatoare în intervale numerice asociate fiecărui calificativ. Intervalele se definesc şi se acceptă de către un grup de specialişti, fiind reprezentative şi asiguratorii în sensul satisfacerii unor criterii aparţinând unor colectivităţi largi.

### 3. Algoritmi de agregare

Odată înfăptuită operaţiunea de omogenizare a tabelelor de date se poate trece la agregarea acestora. Prin agregare se realizează o grupare a datelor după diverse criterii, stabilite în funcţie de scopul analizei. În urma agregării se obţine o ierarhie de la grupuri singulare (care conţin un singur individ) până la grupul complet, care include întreaga colectivitate. Algoritmii de agregare a datelor se împart în două mari categorii: algoritmi ierarhici de clasificare şi algoritmi neierarhici.

Algoritmii ierarhici sunt cei mai utilizaţi în practică. Forma generală a unui algoritm de agregare ierarhic este următoarea:

***Intrări*:** un set de $n(n\text{-}1)/2$ distanţe dintre indivizi

***Pasul* 1**: se determină cea mai mică distanţă

***Pasul* 2:** se adună termenii $i$ şi se înlocuiesc cu un nou obiect $i\cup k$, actualizându-se şi distanţele pentru toţi termenii $j\neq i$, $k$, prin relaţia:

$d_{i\cup k,j} = \min\{d_{ij}, d_{kj}\}$

Se elimină distanţele $d_{ij}$ şi $d_{kj}$ pentru toate valorile $j$ atâta timp cât nu mai sunt folosite.

**Pasul 3:** Atâta timp cât rămân cel puţin doi termeni, se revine la pasul 1.

Distanţele egale pot fi tratate într-o ordine arbitrară. Sunt $n$-1 adunări la pasul 2. Poate fi convenabil să se indexeze grupurile găsite la pasul 2 cu $n$+1, $n$+2,... 2$n$-1, sau o altă variantă este indexarea grupului $i\cup k$ cu cel mai mic indice dintre $i$ şi $k$. Denumirea de *o singură legătură* este dată deoarece la pasul 2 distanta de interconectare dintre două grupuri sau componente ($i\cup k$ şi $j$) este definită ca cea mai mică distanţă de interconectare dintre un membru al unui grup şi un membru al altuia. Alte metode ierarhice sunt caracterizate prin diverse funcţii de distanţă ale legăturilor de interconectare.

Pentru că avem $n$-1 adunări, deci iteraţii, şi pentru că pasul 2 necesită un număr de operaţii mai mici decât $n$, acest algoritm-şablon este de complexitate $0(n^2)$.

***Dendograma*** este cea mai răspândită modalitate grafică de reprezentare a ierarhiei rezultate în urma aplicării unui algoritm de agregare.

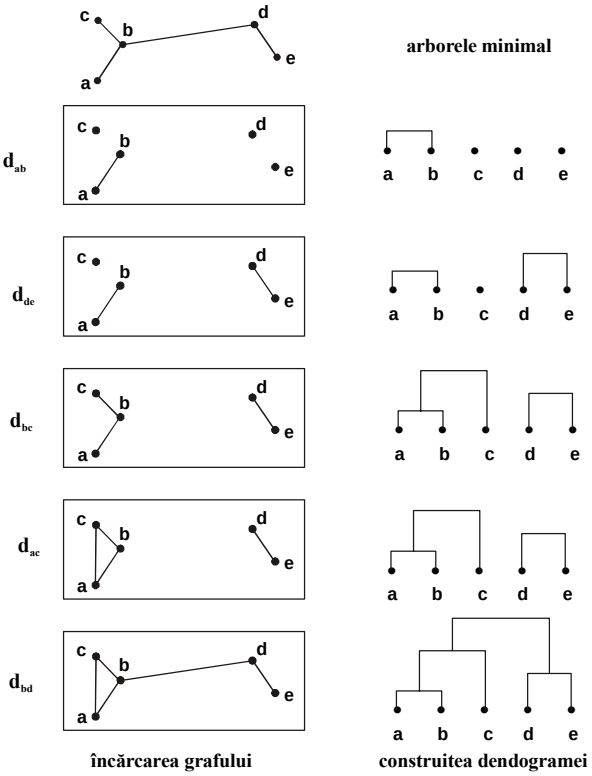



**Fig.1.** Model pentru construirea unei dendograme cu o singură legătură

Există o mare varietate de astfel de algoritmi, de aceea aceştia au fost împărţiţi în două grupe de metode. Prima grupă este cea a me-

todelor cu legături. Acestea sunt metode pentru care reprezentarea cea mai potrivită se face cu ajutorul grafurilor aşa cum se poate observa în figura 1.

Al doilea grup de metode de grupare ierarhice sunt metodele care permit specificarea centrilor grupurilor (ca o medie a membrilor grupului). O metodă convenabilă este cea bazată pe relaţia de actualizare a distanţelor Lance-Williams. Dacă obiectele *i* şi *j* sunt grupate în clusterul $i \cup j$, atunci trebuie doar să fie specificată noua distanţă dintre grup şi toate celelalte puncte (obiecte sau grupuri). Relaţia de actualizare a distanţelor este:

$$d(i \cup j, k) = \alpha_i d(i,k) + \alpha_j d(j,k) + \beta d(i,j) + \gamma |d(i,k) - d(j,k)|$$

unde parametrii $\alpha_i$, $\alpha_j$, $\beta$ definesc criteriul de agregare. Valorile acestora sunt date în coloana a doua a tabelului 8.1. În cazul metodei cu o singură legătură, dacă:

$$\alpha_i = \alpha_j = 1/2,\ \beta = 0 \text{ şi } \gamma = 1/2,$$

atunci

$$d(i \cup j, k) = \frac{1}{2} d(i,k) + \frac{1}{2} d(j,k) - \frac{1}{2} |d(i,k) - d(j,k)|,$$

care poate fi verificată luând ca exemplu 3 puncte $i, j, k$ :

$$d(i \cup j, k) = \min\{d(i,k), d(j,k)\}.$$

Utilizând alte formule de actualizare, pot fi obţinute alte metode implementate într-un mod asemănător, câteva exemple fiind prezentate în tabelul 4.

**Tabelul 4.** Specificaţiile câtorva metode ierarhice de grupare

| Metode ierarhice de grupare | Forma de actualizare a distanţelor | Coordonatele centrului grupului | Distanţe între centre |
|---|---|---|---|
| Legătură simplă | $\alpha_i = 0.5,\ \beta = 0,\ \gamma = -0.5$ $(\min\{d_{ik}, d_{jk}\})$ | | |
| Legătură completă | $\alpha_i = 0.5,\ \beta = 0,\ \gamma = 0.5$ $(\max\{d_{ik}, d_{jk}\})$ | | |
| Media grupului | $\alpha_i = \lvert i\rvert/(\lvert i\rvert+\lvert j\rvert),\ \beta = 0,\ \gamma = 0$ | | |
| Metoda McQuitty (WPGMA) | $\alpha_i = 0.5,\ \beta = 0,\ \gamma = 0$ | | |
| Metoda mediană (WPGMC) | $\alpha_i = 0.5,\ \beta = -0.25,\ \gamma = 0$ | $g = (g_i + g_j)/2$ | $\lVert g_i - g_j\rVert^2$ |
| Centroid (UPGMC) | $\alpha_i = \lvert i\rvert/(\lvert i\rvert+\lvert j\rvert)$ $\beta = -\lvert i\rvert\cdot\lvert j\rvert / (\lvert i\rvert+\lvert j\rvert)^2,\ \gamma = 0$ | $g = \frac{\lvert i\rvert \cdot g_i + \lvert j\rvert \cdot g_j}{\lvert i\rvert + \lvert j\rvert}$ | $\lVert g_i - g_j\rVert^2$ |
| Metoda Ward (variaţie minimă) | $\alpha_i = \frac{\lvert i\rvert + \lvert k\rvert}{\lvert i\rvert + \lvert j\rvert + \lvert k\rvert}$ $\beta = -\frac{\lvert k\rvert}{\lvert i\rvert + \lvert j\rvert + \lvert k\rvert},\ \gamma = 0$ | $g = \frac{\lvert i\rvert \cdot g_i + \lvert j\rvert \cdot g_j}{\lvert i\rvert + \lvert j\rvert}$ | $\frac{\lvert i\rvert \cdot \lvert j\rvert}{\lvert i\rvert + \lvert j\rvert} \lVert g_i - g_j\rVert^2$ |

Menţionăm că:

- $|i|$ este numărul de obiecte în grupul *i*,
- $g_i$ este un vector în spaţiul *m* (*m* este un set de atribute), un punct iniţial sau un centru de grup,
- $\|.\|$ reprezintă norma în metrica Euclidiană,
- denunumirile UPGMA etc. sunt conform Sneath şi Soka,
- formula de recurenţă Lance si Williams este $d_{(i \perp j),k} = \alpha_i d_{ik} + \alpha_j d_{jk} + \beta d_{ij} - \gamma(d_{ik} - d_{jk})$.

În cazul metodelor care utilizează centrul grupului, sunt date coordonatele centrului (coloana 3, tabelul 4) şi distanţele definite între centrele grupurilor (coloana 4, tabel 4). Distanţa euclidiană poate fi utilizată iniţial pentru echivalenţa între două moduri de abordare.

În cazul metodei mediane, se consideră următorul exemplu: fie *a* şi *b* două puncte (vectori *m*-dimensionali reprezentând obiecte sau centre ale grupurilor) care au fost agregate anterior şi fie *c* un alt punct. Din formula de acutalizare a distanţelor Lance-Williams, utilizând distanţa euclidiană pătratică, se obţine:

$$d^2(a \cup b, c) = \frac{d^2(a,c)}{2} + \frac{d^2(b,c)}{2} - \frac{d^2(a,b)}{4} = \frac{\|a-c\|^2}{2} + \frac{\|b-c\|^2}{2} - \frac{\|a-b\|^2}{4}.$$

Noul centru al grupului este $(a+b)/2$, aşa că distanţa sa până la punctul *c* este $\left\| c - \frac{a+b}{2} \right\|^2$.

Pentru metodele care utilizează centrul grupului şi cu modificări potrivite pentru metodele cu grafuri, algoritmul următor este o alternativă la algoritmul bazat pe distanţa generală. Descrierea algoritmului este următoarea:

***Pasul* 1.** Se examinează toate distanţele dintre puncte şi se formează un grup din cele mai apropiate două puncte.

***Pasul* 2.** Se înlocuiesc cele două puncte grupate cu un punct reprezentativ, de exemplu, centrul de greutate.

**Pasul 3:** Se revine la pasul 1, tratând grupurile ca obiecte rămase, până ce obiectele sunt cuprinse într-un singur grup.

La paşii 1 şi 2 noţiunea de *punct* se referă atât la obiecte cât şi la grupuri, amândouă fiind definite ca vectori. Acest algoritm este justificat prin consideraţii de stocare, deoarece numărul de operaţiuni de stocare este de ordinul O(*m*) pentru obiectele iniţiale şi O(*n*) pentru *n*-1 grupuri. În cazul metodelor cu legături, termenul *fragment* se referă la o componentă conectată (metoda cu o singură legătură) sau un subgraf complet (metoda cu legături complete). Complexitatea generală a algoritmului este O($n^3$).

## 4. Aplicaţie

Reluăm exemplul din tabelul 3, cu date privind oferta de maşini de vânzare ale unei societăţi de profil. Tabelul eterogen este următorul:

**Tabelul 5.** Tabelul neomogen

| Cod | *Marca* | *Preţ* | *Consum mediu* | *Garanţie* | *Capacitate Cilindrică* |
|---|---|---|---|---|---|
| C1 | Renault Clio | 8500 | 7 | 3 | 1400 |
| C2 | Dacia 1300 | 150000000 | 8.5 | 18 | 1400 |
| C3 | Citroen C3 | 9000 | moderat | 3 | medie |
| C4 | Audi A80 | 3500 | mic | 100000 | mare |
| C5 | Dacia 1300 | 55000000 | moderat | 0 | 1400 |
| C6 | WW Golf | 200000000 | mare | 150000 | 2000 |
| C7 | Renault Megane | 9000 | 6.5 | 3 | 1600 |
| C8 | Daewo Matiz | 2500 | mic | 15 | 800 |
| C9 | Daewo Cielo | 4000 | moderat | 1 | 1600 |
| C10 | Mercedes | 4000 | mare | 0 | mare |
| C11 | Daewo Tico | 2000 | mic | 6 | mica |

Variabila *Preţ* prezintă două subdomenii definite:

1. $D_1$ = [2000,9000]∩N, pentru preţurile exprimate în Euro,
2. $D_2$ = [55000000,200000000]∩N, pentru preţurile exprimate în lei.

Variabila *Consum mediu* are subdomeniile:

1. $D_1$ = [6.5,8.5], pentru consumul exprimat în litrii de combustibil consumaţi la 100 km parcurşi,
2. $D_2$ = {*mic*,*moderat*,*mare*} pentru calificative de apreciere a consumului.

Variabila *Garanţie* este definită pe subdomeniile:

1. $D_1$ = [0,18], pentru garanţia exprimată în luni,
2. $D_2$ = [0,3], pentru garanţia exprimată în ani,
3. $D_3$ = [0,150000] pentru garanţia exprimată în kilometrii parcurşi.

Pentru variabila *Capacitate cilindrică* subdomeniile sunt:

1. $D_1$ = [800,2000] pentru numărul de $cm^3$ ,
2. $D_2$ = {*mica*, *mare*, *medie*} pentru calificative acordate.

Limitele inferioare şi cele superioare ale domeniilor numerice sunt calculate după datele din tabel. În procesul de omogenizare pot fi folosite ca limite inferioare şi limite superioare valori considerate admisibile pentru domeniul respectiv.

Vom aplica următoarele metode de omogenizare:

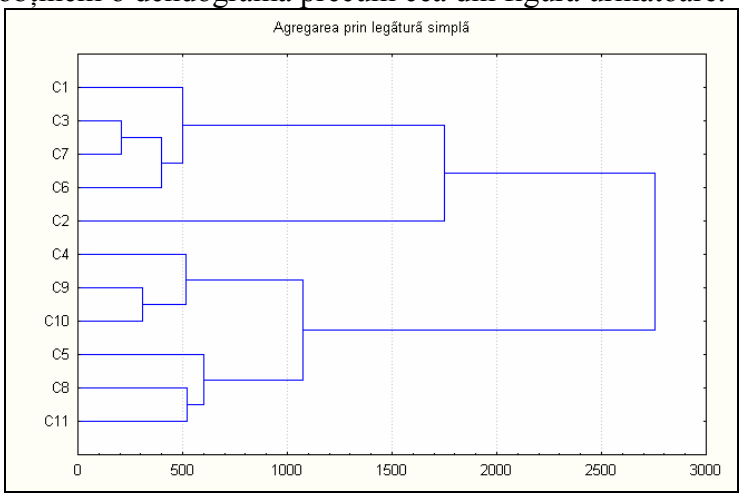
1. omogenizare prin generare de numere pseudoaleatoare pentru domeniul D2 al variabilei *Consum mediu* şi domeniul D2 al variabilei *Capacitate cilindrică*,
2. omogenizare prin normalizare pentru celelalte domenii de tip numeric.

În urma omogenizării obţinem tabelul 6:

**Tabelul 6.** Tabelul omogen

| Cod | *Marca* | *Preţ* | *Consum mediu* | *Garanţie* | *Capacitate Cilindrică* |
|---|---|---|---|---|---|
| C1 | Renault Clio | 8500 | 7 | 3 | 1400 |
| C2 | Dacia 1300 | 6750 | 8.5 | 3 | 1400 |
| C3 | Citroen C3 | 9000 | 6.7 | 3 | 1390 |
| C4 | Audi A80 | 3500 | 4.2 | 2 | 1720 |
| C5 | Dacia 1300 | 2475 | 6.3 | 0 | 1400 |
| C6 | WW Golf | 9000 | 9.4 | 3 | 2000 |
| C7 | Renault Megane | 9000 | 6.5 | 3 | 1600 |
| C8 | Daewo Matiz | 2500 | 3.7 | 2.5 | 800 |
| C9 | Daewo Cielo | 4000 | 7.4 | 1 | 1600 |
| C10 | Mercedes | 4000 | 10.5 | 0 | 1910 |
| C11 | Daewo Tico | 2000 | 4.1 | 1 | 950 |

În urma agregării obţinem o dendogramă precum cea din figura următoare.




Agregarea s-a făcut pe baza distanţei euclidiene simple prin legătură simplă.

**Bibliografie**

1. **Benzecri,J.,P.**, *L'analyse des donnees.La Taxionomie*, Dunod, 1973
2. **Diday, E., Pouget, J., Lemaire, J., Testu, F.**, *Elements d'analyse de donnee*, Dunod, Paris, 1985
3. **Murtagh, F., Heck, A.**, *Multivariate data analysis*, Dordrecht, Holland, 1987
4. **Roux, M.**, *Algorithmes de classification*, Masson, Paris, 1986
5. **Isaic-Maniu, A., Mitruţ, C., Voineagu, V.**, *Statistica pentru managementul afacerilor. Ediţia a doua*, Editura Economică, Bucureşti, 1999
6. **Voineagu, V., Furtună, F.,Voineagu, M, Ştefănescu, C**., *Analiza factorială a fenomenelor social-economice în profil regional*, Editura Aramis, 2002