METODA VODITELJEV V PROGRAMSKIH PAKETIH CLUSE IN SPSS/PC+

Janez Jug

LEADER METHOD IN THE COMPUTER PROGRAMS CLUSE AND SPSS/PC+ - In the paper the leader (k-means) method of nonhierarhic clustering is described and two procedures of it are compared: LEADER implemented in CLUSE, which was developed and is currently running on DECsystem-10 computer in Ljubljana, and QUICK CLUSTER (SPSS/PC+). Both perform k-means clustering (nearest centroid sorting) on a large number of cases into a requested number of groups. They use the squared Euclidean distance and minimize Ward's criterion function. The procedures differ in algorithm and output. LEADER enables to use the constraints (fixed group radius, minimum and maximum number of units in each group), e.g. overlapping or incomplete classification. QUICK CLUSTER can select k cases with well-separated values as initial leaders of clusters and saves the distance from each case to its classification cluster center as new variable on the active file.

METODO DE GVIDANTOJ EN KOMPUTILAJ PROGRAMOJ CLUSE KAJ SPSS/PC+ - En la referaĵo estas priskribita la metodo de gvidantoj (de k-averaĝoj), uzata por nehierarkia grupigado (klasifikado), kaj komparitaj du procedoj de ĝi: LEADER el CLUSE, kiu estis evoluita kaj nun funkcias en komputilo DECsystem-10 en Ljubljana) kaj QUICK CLUSTER el SPSS/PC+. Ambaŭ ebligas k-averaĝan grupigadon (klasifikadon al la plej proksimaj centroidoj) de granda nombro de unuoj en difinitan nombron de grupoj. Ili uzas la kvadrigitan Eŭklidan distancon kaj minimumigas kriterian funkcion de Ward. La procedoj diferencas en algoritmo kaj eligo. LEADER ebligas uzadon de limigoj (fiksita radiuso de grupo, minimuma kaj maksimuma nombro de unuoj en ĉiu grupo), ekz. interkovradon aŭ nekompletan grupigon. QUICK CLUSTER povas elekti k unuojn kun bone apartigitaj valoroj por komencaj gvidantoj de grupoj kaj enigas distancon inter ĉiu unuo kaj ĝia klasifika grupocentro kiel novan variablon en aktivan rikordaron.

Program LEADER iz programskega paketa za razvrščanje v skupine CLUSE (Batagelj, 1984)[1] in podprogram QUICK CLUSTER iz statističnega programskega paketa SPSS/PC+™ temeljita na metodi, ki je v literaturi znana kot metoda voditeljev (leader method), metoda k-povprečij (k-means) ali razvrščanje k najbližjim

---

1. Batagelj ga je začel razvijati leta 1976 in ga še vedno dopolnjuje Programski paket CLUSE deluje na računalniku DECsystem-10 Računalniškeg centra Univerze Edvarda Kardelja v Ljubljani.

središčem (nearest centroid sorting)[2]. Uporabljamo ju lahko za razvrščanje ali uvrščanje enot v skupine[3]. Z obema lahko analiziramo večje množice (nekaj tisoč) enot[4], ki so podane kot realni vektorji na vhodni datoteki.

Vsaka enota $X_i \in E$ (množica enot) je podana z vektorjem:

$$X_i = (x_{i1}, x_{i2}, \ldots, x_{im}) \quad i = 1,..,n; \; j = 1,..,m$$

kjer je n število enot, m število spremenljivk, x_{ij} pa vrednost j-te spremenljivke za i-to enoto. Spremenljivke morajo biti razmernostne, intervalne ali takšne ordinalne, da jih lahko imamo za intervalne. Če so merjene v različnih merskih enotah, jih moramo standardizirati. Za take enote lahko postavimo za voditelje skupin kar njihova središča. Začetne voditelje lahko podamo ali pa jih program sam določi. Voditelje določimo na podlagi teoretičnih domnev ali rezultatov predhodnih raziskav.

1. Metoda voditeljev

Osnovna zamisel metode voditeljev je naslednja: v množici enot izberemo k voditeljev (predstavnikov, središč, pogosto kar težišč skupin), enote pa pridružimo najbližjim voditeljem. Nato razvrstitev iterativno popravljamo, dokler se voditelji ne ustalijo.

Osnovna shema metode voditeljev (Batagelj, 1985, 85) je:[5]

določi C_0, $C \leftarrow C_0$
ponavljaj
  za vse $C \in C$ določi voditelja L
  novo razvrstitev C dobiš tako, da prirediš vsako
    enoto njej najbližjemu voditelju
dokler se voditelji ne ustalijo

---

2. Središče ali težišče (tudi centroid ali center) je za numerične podatke določeno s povprečnimi vrednostmi za vsako od m spremenljivk. Izračunamo ga takole:

$$T_i = (1/n) \sum_{j=1}^{m} X_{ij}$$

3. Pri uvrščanju (klasifikaciji) so skupine in značilnosti skupin že določene, zato je treba vsako enoto prirediti njej najbližjemu voditelju skupine, pri razvrščanju pa je treba odkriti skupine in ugotoviti njihove značilnosti.

4. Enoto dobimo, ko na vsakem izmed n objektov (anketirancev, itd.) izmerimo m lastnosti (karakteristik), ki glede na naše predpostavke in namene objekt dovolj natančno opisujejo. (Ferligoj, 1980, 13.)

5. Namesto velikih pisanih črk uporabljam kurzivne velike tiskane črke, namesto zavitih oklepajev znaka $\{\ \}$, kot oznake za absolutno vrednost pa $|\ |$.

Postopek razvrščanja se začne z izbiro k točk v m-razsežnem prostoru, ki so začetni voditelji[6]. Metoda voditeljev dopušča prestavitve enot, tako da slabe začetne razvrstitve lahko popravimo v naslednjih fazah. Metoda razvrsti enote v skupine tako, da optimizira kriterijsko funkcijo[7]. Kadar si lahko enote predstavimo v m-razsežnem prostoru, se pogosto opremo na geometrijsko-fizikalno sliko in si skupine predstavljamo kot nekakšne oblake. Tedaj je zelo priljubljena mera za napako razvrstitve Wardova kriterijska funkcija (Batagelj, 1985, 40).

Pri metodi voditeljev ima Wardova kriterijska funkcija obliko

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} p(C) \qquad \text{in} \qquad p(C) = \sum_{X \in C} d(X,L)$$

pri čemer je d različnost (navadno kvadrat evklidske razdalje) in L voditelj oz. središče skupine C[8]. Wardovo kriterijsko funkcijo zapišemo tudi kot

$$P(\mathcal{C}) = sl(W) \qquad \text{in} \qquad W = \sum_{i=1}^{k} W_i$$

kjer je W matrika razpršenosti (disperzije) znotraj skupin, W_i pa matrika razpršenosti za i-to skupino. (Batagelj, 1985, 22; Everitt, 1974, 24 in 26.)

Razvrstitev je lahko odvisna od izbire začetnih voditeljev in začetne urejenosti enot. Ko se postopek konča, tj., ko nobena prestavitev enote ne izboljša razvrstitve, je dosežen lokalni minimum kriterijske funkcije. Dobljeno razvrstitev lahko sprejmemo kot rešitev ali pa, podobno kot pri vseh postopkih lokalne optimizacije, poskušamo dobiti 'dobro' rešitev in vtis o njeni kakovosti tako, da postopek večkrat ponovimo z različnimi začetnimi voditelji (Batagelj, 1985, 85). V splošnem ni načina, kako izvedeti, ali je bil dosežen absolutni ali lokalni minimum kriterijske funkcije (Everitt, 1974, 26). Eden največjih problemov tehnik razvrščanja, ki poskušajo optimizirati kak kriterij, je, da so podoptimalne rešitve pogoste (Everitt, 1974, 62).

---

6. Za izbiro teh točk je bilo predlaganih več postopkov. O tem glej: Anderberg (1973, 157-159).

7. Kriterijska funkcija $P(\mathcal{C})$ vsaki razvrstitvi R priredi neko pozitivno realno število $P(\mathcal{C}) : \mathcal{C} \to R^+$.

8. V francoski literaturi količino $P(\mathcal{C})$ (vsota kvadratov odstopanj) pogosto imenujejo inercija ali vztrajnost skupine C in jo označujejo z I(C).

Z metodo voditeljev lahko dobimo popolno razvrstitev (razbitje) množice enot v k skupin (vsako enoto razvrstimo v skupino z njej najbližjim voditeljem) ali pa razvrstitev v k ne nujno ločenih skupin, kjer tudi ni nujno, da razvrstimo vse enote. Primer take metode voditeljev je naslednji: dana je množica k voditeljev in realno število r > 0. Pripadnost enote X posamezni skupini C_i je določena s pogojem

$X \in C_i$ natanko takrat, ko je $d(X,L_i) \leq r$

kar pomeni, da enota, ki leži v okolici i-tega in j-tega voditelja, je element i-te in j-te skupine. Enota, ki ni v okolici nobenega voditelja, ni razvrščena (Ferligoj, 1982, 55).

Pri razvrščanju v skupine uporabljamo različne mere podobnosti ali različnosti med enotami ali spremenljivkami. Na posameznih področjih naletimo na najrazličnejše mere različnosti. Tiste različnosti, ki zadoščajo naslednjim pogojem:

a.	$d(X,Y) \geq 0$	nenegativnost
b.	$d(X,X) = 0$	
c.	$d(X,Y) = d(Y,X)$	simetričnost
č.	$d(X,Y) = 0 \Rightarrow X = Y$	razločljivost
d.	$\forall Z: d(X,Y) \leq d(X,Z) + d(Z,Y)$	trikotniška neenakost

so razdalje (Ferligoj, 1982, 8-9). Razdalje Minkowskega med enotama X in Y so definirane takole:

$$d_p(X,Y) = \left(\sum_{j=1}^{m} |x_j - y_j|^p \right)^{1/p}$$

Pri p = 1 dobimo razdaljo Manhattan (ali city-block), pri p = 2 evklidsko razdaljo, ko je p neskončno velik, pa trdnjavsko ali razdaljo Čebiševa (Batagelj, 1985, 28).

2. Program LEADER

Program LEADER temelji na metodi voditeljev, ki je poseben primer metode (dinamičnih) oblakov (Ferligoj, 1983, 92). Od 'klasične' metode voditeljev (Hartigan, 1975; Diday, 1979)[9] se razlikuje po tem, da omogoča tudi določanje razvrstitev, ki zadoščajo dodatnim omejitvam.

---

9. Navedeno po: Ferligoj, 1983.

Na delovanje programa LEADER lahko vplivamo z naslednjimi omejitvami (parametri) (Ferligoj, 1983, 92):
- največje število skupin / voditeljev (maxldr)[10];
- polmer (r) krogle okrog voditelja, v kateri morajo biti vse enote skupine[11];
- največje število enot v posamezni skupini (maxuni);
- dovoljeno povečanje polmera r, če je dosežen maxldr in je od vsakega voditelja enota oddaljena za več kot r;
- pri izbiri novih voditeljev (središč skupin) lahko zahtevamo, da veljajo samo tisti, za katere vsebujejo pripadajoče skupine dovolj (vsaj minuni) enot.

Če uporabimo te omejitve, potem ne dobimo vedno popolnih razvrstitev v k skupin. Nekatere enote lahko ostanejo nerazvrščene (skupina 0).

Program LEADER ima štiri mere različnosti d med enotami:

1. kvadrat evklidske razdalje

$$d_2^2(X,Y) = \sum_{j=1}^{m} (x_j - y_j)^2$$

2. razdaljo Manhattan

$$d_1(X,Y) = \sum_{j=1}^{m} |x_j - y_j|$$

3. razdaljo Čebiševa

$$d_\infty(X,Y) = \max_{j=1}^{m} |x_j - y_j|$$

4. razdaljo Canberra (za pozitivne vektorje)

$$d_{CANB}(X,Y) = \sum_{i=1}^{m} |(x_j - y_j)/(x_j + y_j)|$$

---

10. Lahko zahtevamo največ 2856 voditeljev.

11. Pripadnost enote $X \in E$ posamezni skupini C_i je določena s pogojem:

$$X \in C_i \iff d(X,L_i) \leq r$$

To omogoča delno uvrstitev (klasifikacijo) enot v največ k, ne nujno ločenih skupin. Če določimo prevelik polmer krogle, lahko kakšna enota pade hkrati v dve skupini. Če hočemo dobiti popolno razvrstitev, določimo polmer r = 0 in povečanje polmera po potrebi.

Kriterijska funkcija, ki jo LEADER poskuša minimizirati, je

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} \sum_{X \in C} d(X,L)$$

Zaradi možnosti nepopolnih razvrstitev izpisuje tri vrednosti, povezane s kriterijsko funkcijo:

$$Pc = P(\mathcal{C}_c)$$

$$Pt = P(\mathcal{C}_t)$$

$$Pe = \sum_{X \in C_\bullet} \min_{L \in \mathcal{L}} d(X,L)$$

pri čemer razvrstitev $\mathcal{C}_c$ sestavljajo skupine, ki zadoščajo vsem omejitvam, razvrstitev $\mathcal{C}_t$ skupine, ki vsebujejo premalo enot, $C_\bullet$ je skupina izločenih enot, $\mathcal{L}$ pa množica voditeljev v tekočem koraku. Tako je npr. vrednost razvrstitve, ki jo (lahko) shranimo na datoteko za nadaljnjo analizo, enaka $P_c + P_t$.

Delovanje programa krmilimo s kontrolno spremenljivko (cont). Njene vrednosti imajo naslednji pomen:

0 - shrani rezultate in končaj
1 - opravi še en korak
2 - spremeni parametre + (1)
3 - preberi voditelje + (2)

Postopek programa LEADER je naslednji (Ferligoj, 1983, 95):

```
step <- 0; cont <- 4; save <- 0; extend <- false; numldr <- 0;
ponavljaj
     step <- step + 1
     če je cont ≥ 2 potem spremeni parametre: maxldr, maxuni, r,
          extend, save
     če je cont ≥ 3 potem preberi voditelje
     numrow <- 0; nexcld <- 0; Pe <- 0;
     (rmax_i, p_i, n_i) <- 0, i = 1,...,maxldr
     ponavljaj, dokler ne zmanjka podatkov
          preberi tekočo enoto; numrow <- numrow + 1;
          i <- argmin (L_k, enota); d <- d(L_i, enota);
                   k
          če je (d ≤ r)∧(n_i ≤ maxuni) potem
               dodaj enoto skupini i:
               n_i <- n_i + 1; p_i <- p_i + d;
               rmax_i <- max(d, rmax_i)
          sicer, če je numldr < maxldr potem
               proglasi enoto za novega voditelja:
               numldr <- numldr + 1; L <- L U {enota}
          sicer, če je extend, potem
               popravi polmer in dodaj enoto skupini i:
               r <- d; n_i <- n_i + 1; p_i <- p_i + d; rmax_i <- d
          sicer
```

```
                enote ni mogoče razvrstiti:
                nexcld <- nexcld + 1; Pe <- Pe + d
            izpiši i na pomožno datoteko;
        Pc <- 0; Pt <- 0; error <- 0; delmax <- 0; rmax <- 0;
        preberi minuni
        za i = 1,2,...,numldr  ponovi
            določi težišče skupine i in pomik delta glede na
            voditelja
            če je n_i ≥ minuni,  potem
                proglasi težišče za voditelja:
                delmax <-  max(delmax,delta);  error  <-  error  +
                delta;
                rmax <- max(rmax,rmax_i); Pc <- Pc + p_i
            sicer
                Pt <- Pt + p_i
            izpiši podatke o skupini i
        izpiši podatke o razvrstitvi;
        na zahtevo shrani razvrstitev in/ali težišča;
        preberi cont - odloči se o naslednji akciji
dokler ni  cont ≤ 0;
```

Program LEADER izpisuje sporočila o poteku analize na izhodno datoteko "ime.LST" in pomembnejša tudi na zaslon. Na uporabnikovo zahtevo shrani na datoteki "ime.LDN" središča posameznih skupin in na datoteki "ime.ANA" razvrstitev enot. Izpis središč in razvrstitve lahko zahtevamo v vsakem koraku (iteraciji), navadno pa to storimo, ko se postopek konča.

Na datoteko "ime.LST" najprej izpiše določene omejitve (maxldr, r, maxuni) in število prebranih voditeljev. Za vsako iteracijo posebej izpiše število prebranih enot, število neuvrščenih enot, število voditeljev, najmanjše število enot v skupini, določeno z omejitvijo 'minuni', osnovne podatke o posamezni skupini, tri vrednosti kriterijske funkcije, največji premik voditelja, napako (vsoto vseh premikov voditeljev) ter število shranjenih voditeljev in število nepokritih enot. Osnovni podatki o skupini vsebujejo število enot, polmer (razdaljo med voditeljem in najoddaljenejšo enoto v skupini v kvadratu evklidske razdalje), vrednost, ki jo skupina prispeva h kriterijski funkciji, in premik voditelja (kvadrat razlike med vrednostima prve spremenljivke tekočega voditelja in središča skupine).

3. Podprogram QUICK CLUSTER

Podprogram QUICK CLUSTER je primeren za razvrščanje večjega števila enot v dano število skupin. Enote morajo biti opisane z vsaj ordinalnimi spremenljivkami, ki jih moramo standardizirati, če so merjene v različnih merskih enotah.

Postopek za določanje skupin v programu QUICK CLUSTER temelji na razvrščanju k najbližjim središčem (Anderberg, 1973,

160-167). Enota je razvrščena v tisto skupino, za katero je razdalja med enoto in voditeljem (središčem) skupine najmanjša. Dejanski potek postopka je odvisen od zahtev v ukazu.

Postopek QUICK CLUSTER je razdeljen na tri korake. V prvem izbere začetne voditelje skupin, v drugem izračuna klasifikacijske voditelje, v tretjem koraku pa razvrsti enote h klasifikacijskim voditeljem in izračuna končna središča skupin.

Program QUICK CLUSTER izbere začetne voditelje na tri načine: (1) prebere tiste, ki jih poda uporabnik, (2) vzame prvih k enot brez manjkajočih vrednosti, (3) izmed vseh enot izbere k enot brez manjkajočih vrednosti, ki imajo velike razdalje med seboj. Izbiranje enot za začetne voditelje začne tako, da prvih k enot v datoteki, pri čemer je k število zahtevanih skupin, izbere kot začasne voditelje. Nato pregleda naslednje enote. Če je najmanjša razdalja med enoto in najbližjim voditeljem večja od razdalje med dvema najbližjima voditeljema, potem ta enota nadomesti najbližjega voditelja. "Enota torej nadomesti središče. če je najmanjša razdalja od enote do središča večja od najmanjše razdalje med tistim središčem in vsemi drugimi središči." (Norušis, 1986, B-95 - B-96).

Potem ko določi začetne voditelje, nadaljuje takole: Za vsako enoto izračuna kvadrat evklidske razdalje[12] med enoto in vsemi voditelji. Nato enoto pridruži najbližjemu voditelju. Vsakokrat, ko doda enoto v skupino, izračuna novo središče te skupine, in sicer "k (novi) povprečni vrednosti vseh enot v skupini plus začetno središče skupine" (Norušis, C-84) tj. tako, da pri računanju središča upošteva vse enote, ki so v skupini, in začetno središče skupine (SPSS, 1986, 171). Potem ko so vse enote razvrščene, izračuna povprečne vrednosti za spremenljivke iz enot, ki so bile dodeljene vsaki skupini, in enot, ki so bile začetni voditelji skupin. Te vrednosti so klasifikacijski voditelji (classification centers) skupin. Kadar izpustimo ta korak, vzame začetne voditelje skupin kot klasifikacijske in takoj začne tretji korak.

V tretjem koraku vse enote ponovno razvrsti v skupine z najbližjim klasifikacijskim voditeljem in izračuna končna središča skupin. Pri računanju končnih središč skupin klasifikacijskih voditeljev ne upošteva.

Programski paket SPSS/PC+ izpiše rezultate v datoteko SPSS.LIS ali v datoteko, določeno z ukazom SET LISTING='ime'. Podprogram QUICK CLUSTER izpiše na ekran in/ali na izhodno datoteko: začetne voditelje skupin, klasifikacijske voditelje skupin, pripadnost vsake enote skupini, razdaljo med enoto in klasifikacijskim voditeljem skupine, končna središča skupin, razdalje med končnimi središči skupin, analizo variance za vse

---

12. "V tem postopku bi tudi nekvadrirana evklidska razdalja dodelila enote na enak način." (Norušis, 1986, C-85)

spremenljivke in število enot v vsaki skupini. Lahko zahtevamo, da pripadnost vsake enote skupini in razdaljo med enoto in klasifikacijskim voditeljem skupine shrani v delujočo datoteko. Če to storimo, lahko z drugimi postopki iz razdalje med enoto in klasifikacijskim voditeljem skupine izračunamo še vrednost Wardove kriterijske funkcije ter najmanjšo in največjo razdaljo med enoto in voditeljem v vsaki skupini.[13]

4. Primerjava

Program LEADER in podprogram QUICK CLUSTER se razlikujeta v načinu obdelave, v obravnavanju manjkajočih vrednosti, v postopku, v določanju začetnih voditeljev, v številu vgrajenih razdalj in v izpisu rezultatov. Najpomembnejša razlika med njima pa je, da LEADER omogoča razvrščanje z omejitvami, QUICK CLUSTER pa ne.

Obdelava s programom LEADER je interaktivna na računalniku DECsystem-10, s podprogramom QUICK CLUSTER pa paketna na osebnem računalniku. Pri programu LEADER pridemo do končne rešitve v eni obdelavi z več koraki (iteracijami). S podprogramom QUICK CLUSTER moramo narediti več obdelav, da dobimo dobro rešitev. Nadaljujemo tako, da v drugi in naslednjih obdelavah podamo kot začetne voditelje končna središča iz prejšnje obdelave in izpustimo drugi korak postopka. S tem dosežemo, da QUICK CLUSTER deluje enako kot LEADER brez omejitev.

Program LEADER zahteva enote brez manjkajočih vrednosti, QUICK CLUSTER pa ne, ker ima vgrajene tri načine ravnanja z manjkajočimi vrednostmi[14].

Programa se razlikujeta v postopku razvrščanja (Glej sliko 1!). QUICK CLUSTER lahko izbere za začetne voditelje razen prvih k enot, kakor LEADER, tudi k enot, ki so dobro ločene v prostoru. S takšnimi voditelji pa hitreje pridemo do manjše vrednosti

---

13. Te rezultate dobimo, če za stavkom QUICK CLUSTER vstavimo:

```
COMPUTE       KF=RAZDALJA**2.
MEANS         TABLES=KF  BY SKUPINA/OPTIONS=6,7,11.
PROCESS IF    (SKUPINA EQ 1).
DESCRIPTIVES  VARIABLES RAZDALJA/STATISTICS=12,13.
PROCESS IF    (SKUPINA EQ 2).
DESCRIPTIVES  VARIABLES RAZDALJA/STATISTICS=12,13.
```

14. Manjkajoče vrednosti obravnava takole: (1) ne upošteva enot z manjkajočimi vrednostmi, (2) pri računanju ne upošteva spremenljivk z manjkajočimi vrednostmi na posameznih enotah, (3) vključi manjkajoče vrednosti.

Slika 1: Primerjava postopkov

LEADER	QUICK CLUSTER
1. Izbor začetnih voditeljev	1. Izbor začetnih voditeljev
Voditelji so podani. Izbere prvih k enot.	Voditelji so podani. Izbere prvih k enot. Izbere dobro ločene enote.
2. Razvrstitev enot in izračun središč skupin	2. Določanje klasifikacijskih voditeljev
Enote razvrsti k začetnim voditeljem, izračuna središča skupin in jih proglasi za nove voditelje.	Dodaja enote in popravlja središča skupin. Ko porabi vse enote, središča postanejo klasifikacijski voditelji.
Razvrščanje nadaljuje iterativno, dokler se voditelji ne ustalijo.	3. Končna razvrstitev
	Enote razvrsti k najbližjim klasifikacijskim voditeljem in izračuna končna središča.

kriterijske funkcije[15]. Pri programu LEADER se drugi korak ponavlja, dokler se voditelji ne ustalijo (tj., dokler niso v neki ponovitvi središča enaka voditeljem), postopek QUICK CLUSTER pa v drugem koraku določi klasifikacijske voditelje, v tretjem koraku pa enote razvrsti k najbližjim klasifikacijskim voditeljem in izračuna končna središča skupin, ki niso nujno enaka klasifikacijskim voditeljem. Če hočemo doseči, da bodo končna središča enaka klasifikacijskim voditeljem, moramo obdelavo

15. Primerjava vrednosti Wardove kriterijske funkcije

LEADER		QUICK CLUSTER		
Iteracija		Obdelava	prvih k enot	k izbranih enot
1	59 251			
		1	45 203	43 219
2	31 927			
		2	31 779	27 525
3	30 033			

Pri programu LEADER se kriterijska funkcija nanaša na začetne voditelje vsake iteracije, pri postopku QUICK CLUSTER pa na klasifikacijske voditelje.

nadaljevati. V naslednji obdelavi vzamemo končna težišča prve obdelave za začetne voditelje druge obdelave in izpustimo drugi korak postopka (določanje klasifikacijskih voditeljev). To ponavljamo, dokler se končna središča ne razlikujejo več od začetnih voditeljev, ki so hkrati klasifikacijski.

Pri programu LEADER lahko zahtevamo razvrstitev od 1 do največ 2856 skupin, pri podprogramu QUICK CLUSTER pa v 2 ali več skupin. Če pri podprogramu QUICK CLUSTER zahtevamo razvrstitev v eno skupino, razvrsti enote v dve skupini. Ta pomanjkljivost ni tako huda, saj v eno skupino razvrščamo samo zato, da izvemo vrednost kriterijske funkcije. To pa lahko pri podprogramu QUICK CLUSTER izračunamo tako, da podamo dva voditelja. Prvi ima za vrednosti aritmetične sredine spremenljivk, za drugega pa izberemo dovolj velike vrednosti, da bodo vse enote bližje prvemu voditelju. Po razvrstitvi v dve skupini bo druga skupina prazna.

LEADER izpisuje na ekran in na izhodno datoteko vrednost kriterijske funkcije, v podprogramu QUICK CLUSTER pa kriterijska funkcija ni nikjer omenjena. Programa ne minimizirata Wardove kriterijske funkcije, vendar pa jo lahko pri obeh uporabljamo, kajti vsak postopek razvrščanja, ki prestavlja enote iz skupine v skupino, dokler niso v skupini bližje (v evklidskem smislu) svojemu središču, kot pa središču katere koli druge skupine, lahko imamo za poskus minimiziranja sledi matrike razpršenosti znotraj skupin (Everitt, 1974, 26-27) oz. Wardove kriterijske funkcije. Diday je dokazal, da način razvrščanja, ki je vgrajen v programu LEADER, konvergira po Wardovi kriterijski funkciji, tj. zagotavlja minimum te funkcije.

QUICK CLUSTER ima vgrajen samo kvadrat evklidske razdalje, LEADER pa poleg tega še razdaljo Canberra, razdaljo Manhattan in razdaljo Čebiševa, toda le za kvadrat evklidske razdalje velja, da postopek konvergira.

Za prikaz delovanja obeh programov sem izbral podatke, ki jih je Everitt (Everitt, 1974, 29) uporabil za prikaz enega od načinov razbitja množice enot v skupine. Na sliki 2 je vhodna datoteka, na slikah 3, 4 in 5 pa so rezultati obdelav.

Slika 2: Vhodna datoteka

	X_1	X_2
E_1	1.0	1.0
E_2	1.5	2.0
E_3	3.0	4.0
E_4	5.0	7.0
E_5	3.5	5.0
E_6	4.5	5.0
E_7	3.5	4.5

Slika 3: Rezultati programa LEADER (1. iteracija)

Spremenljivki	Začetna voditelja		Končna voditelja	
	L_1	L_2	L_1	L_2
x_1	1.000000	1.500000	1.000000	3.500000
x_2	1.000000	2.000000	1.600000	4.583333
n_i	1	6		
$r_{i,max}$	0.00000	37.25000		
$P(C_i)$	0.00000	84.75000		
premik	0.00000	4.00000		
$P(C)$	84.750			
napaka	4.00000			

n_i = število enot v skupini
$r_{i,max}$ = največja razdalja med voditeljem L_i in enotami v skupini C_i (v kvadratu evklidske razdalje)
$P(C_i)$ = vrednost skupine, tj. vsota kvadratov evklidskih razdalj med vsako enoto in voditeljem skupine
$P(C)$ = vrednost kriterijske funkcije
premik = premik voditelja, tj. kvadrat razlike med vrednostma prve spremenljivke tekočega voditelja in središča skupine
napaka = vsota vseh premikov voditeljev

Slika 4: Rezultati podprograma QUICK CLUSTER (brez 2. koraka)

Spremenljivki	Začetna voditelja		Klasifikacijska voditelja		Končna voditelja	
	L_1	L_2	L_1	L_2	L_1	L_2
x_1	1.0000	1.5000	1.0000	1.5000	1.0000	3.5000
x_2	1.0000	2.0000	1.0000	2.0000	1.0000	4.5833
n_i			1	6		
$r_{i,min}$			0.0	0.0		
$r_{i,max}$			0.0	6.10328		
$P(C_i)$			0.0	84.750		
$P(C)$			84.750			
$d_2(L_1,L_2)$					4.3692	

$r_{i,min}$ in $r_{i,max}$ = najmanjša in največja razdalja med voditeljem L_i in enotami v skupini C_i (v evklidski razdalji)
$d_2(L_1,L_2)$ = evklidska razdalja med končnima voditeljema

Slika 5: Rezultati podprograma QUICK CLUSTER

Spremenljivki	Začetna voditelja		Klasifikacijska voditelja		Končna voditelja	
	L_1	L_2	L_1	L_2	L_1	L_2
x_1	1.0000	5.0000	1.4375	4.4661	1.2500	3.9000
x_2	1.0000	7.0000	1.6000	6.1635	1.5000	5.1000
n_i			2	5	2	5
$r_{i,min}$			0.405	0.992	0.559	0.412
$r_{i,max}$			0.743	2.613	0.559	2.195
$P(C_i)$			0.715	15.158	0.625	7.900
$P(C)$			15.873		8.525	
$d_2(L_1,L_2)$					4.4702	

n_i = število enot v skupini
f_i = odstotek enot v skupini
$r_{i,min}$ = najmanjša razdalja med voditeljem L_i in enotami v skupini C_i (v evklidski razdalji)
$r_{i,max}$ = največja razdalja med voditeljem L_i in enotami v skupini C_i (v evklidski razdalji)
$P(C_i)$ = vrednost skupine, tj. vsota kvadratov evklidskih razdalj med vsako enoto in voditeljem skupine
$P(C)$ = vrednost kriterijske funkcije
$d_2(L_1,L_2)$ = evklidska razdalja med končnima voditeljema

5. Sklep

Prednosti programa LEADER pred podprogramom QUICK CLUSTER so v postopku razvrščanja in izpisu rezultatov. LEADER avtomatično ponavlja korake, dokler ne pride do rešitve, omogoča omejitve in ima več mer različnosti. Med rezultati izpisuje tudi vrednosti kriterijske funkcije, največji polmer in napako, ki nam pomagajo ugotoviti, kdaj se postopek konča. QUICK CLUSTER je glede postopka boljši, ker omogoča ravnanje z manjkajočimi vrednostmi in ker ima vgrajeno izbiranje začetnih voditeljev. V izpisu rezultatov pa v nasprotju s programom LEADER izračuna razdalje med končnimi središči in analizo variance[16]. Pripadnost skupini in razdaljo med enoto in voditeljem lahko shrani kot novi spremenljivki v delujoči datoteki, pri programu LEADER pa je razvrstitev enot v posebni datoteki. Pri podprogramu QUICK CLUSTER lahko z uporabo spremenljivke razdalja izračunamo kriterijsko funkcijo, najmanjši in največji polmer in dosežemo podoben način omejitve kot pri programu LEADER z omejitvijo maxldr.

---

16. V paketu CLUSE lahko izračunamo razdaljo med končnimi težišči, analizo variance pa napravimo s SPSS, če jo potrebujemo.

LITERATURA

Anderberg, M. R. (1973): Cluster Analysis for applications. Academic press, New York.

Batagelj, V. (1984): CLUSE - programi za razvrščanje v skupine. Priročnik za DEC-10. 1. seminar sekcije za uporabno matematiko. DMFA SRS, Ljubljana. Str. 13-20.

Batagelj, V. (1985): Razvrščanje v skupine - nehierarhični postopki. Magistrsko delo. FE, Ljubljana.

Diday, E. (1979): Optimisation en classification automatique. Tome 1., 2. INRIA, Rocquencourt.

Everitt, B. (1974): Cluster analysis. Heinemann Educational Books, London.

Ferligoj, A. (1982): Razvrščanje v skupine. Zapiski. Ljubljana.

Ferligoj, A. (1983): Razvrščanje v skupine - izbrane teme. RI FSPN, Ljubljana.

Hartigan, J. A. (1975): Clustering Algorithms. John Wiley, New York.

Norušis, M. J. (1986): Advanced statistics SPSS/PC+$_{TM}$ For the IBM PC/XT/AT. SPSS Inc, Chicago.

SPSS Statistical Algorithms (1986): 2nd printing revised. SPSS Inc, Chicago.