Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Simona Korenjak - Černe

RAZVRŠČANJE VELIKE MNOŽICE NEENOVITO OPISANIH PODATKOV

DOKTORSKA DISERTACIJA

Mentor: Prof. dr. Ivan Bratko
Somentor: Prof. dr. Vladimir Batagelj

Ljubljana, 2003

Vsebina

Predstavitev problema

Razvrščanje v skupine pomeni formalizirano, načrtovano, namensko ali znanstveno iskanje skupin podatkov (Hartigan, 1975).

Razvrščanje (angl. clustering)

1. **neenovito opisanih (mešanih) podatkov**

   večina znanih hierarhičnih metod razvrščanja temelji na matriki različnosti in so zato zahtevnosti vsaj $\mathcal{O}(n^3)$, npr. Wardova metoda

2. **velikih množic**

   postopki DM so večinoma primerni le za številske podatke,
   npr. BIRCH, DBSCAN, CURE, ScaleKM

V disertaciji razvite rešitve

- preoblikovanje podatkov v enovit opis: uvedba **simbolnega opisa**;
- **zastavitev optimizacijskega problema**: vpeljava različnosti, prilagojenih simbolnim opisom;
- prilagoditev metode voditeljev: teoretično izpeljane **optimalne rešitve določitve predstavnika skupine** za vpeljani različnosti;
- prilagoditev hierarhične metode združevanja: **izbira računanja nivoja skupine, prilagoditev Wardove metode**;
- prilagoditev hierarhične metode dodajanja

Oznake

$\mathbf{X}$	enota,
$\boldsymbol{X}^o$	$\boldsymbol{X}^o = [x_1, x_2, \ldots x_m], x_j = V^j(\mathbf{X})$, vektorski opis enote $\mathbf{X}$,
$\boldsymbol{X}$	simbolni opis enote $\mathbf{X}$,
$\mathbf{E}$	*končna* množica enot, ki jih razvrščamo,
d	$d : \mathbf{E} \times \mathbf{E} \to \mathbb{R}$, *mera različnosti* (angl. dissimilarity)
C	$\emptyset \neq C \subseteq \mathbf{E}$, *skupina* (angl. cluster) enot,
$p(C)$	*napaka skupine* C,
$\mathcal{C}$	$\mathcal{C} = \{C_i\}$, *razvrstitev*,
Φ	*množica dopustnih razvrstitev*,
P	$P : \Phi \to \mathbb{R}_0^+$, *kriterijska funkcija*.

Simbolni opisi enot in skupin - oznake

$\{V_i,\ i = 1, \ldots k_V\}$ razbitje zaloge vrednosti spremenljivke V

$$
\begin{aligned}
Q(i, C; V) &= \{\mathbf{X} \in C : V(\mathbf{X}) \in V_i\},\ i = 1, \ldots, k_V \\
q(i, C; V) &= \mathrm{card}(Q(i, C; V)) \qquad \text{(frekvenca)} \\
f(i, C; V) &= \frac{q(i, C; V)}{\mathrm{card}(C)} \qquad \text{(relativna frekvenca)}
\end{aligned}
$$

Simbolni opis skupine

$$\boldsymbol{C} = [\boldsymbol{C}(V^1), \ldots, \boldsymbol{C}(V^m)],$$

$$\boldsymbol{C}(V^j) = [f(1, C; V^j), \ldots, f(k_j, C; V^j)]$$

Lastnosti simbolnega opisa s porazdelitvami

- **prostor predstavitve posamezne enote ali skupine je neodvisen od števila enot**;
- **usklajenost z združevanjem**:
  za ločeni skupini C_1 in C_2 velja

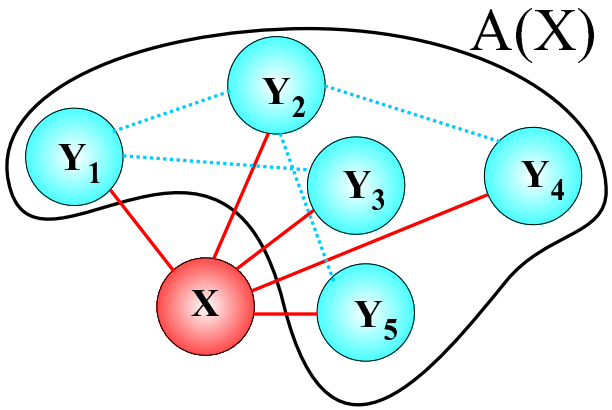
$$f(i, C_1 \cup C_2; V) = \frac{\mathrm{card}(C_1)\, f(i, C_1; V) + \mathrm{card}(C_2)\, f(i, C_2; V)}{\mathrm{card}(C_1) + \mathrm{card}(C_2)};$$

- omogoča **enovit opis za vse vrste spremenljivk**;
- omogoča **enovit opis enot, skupin in predstavnikov skupin**;
- omogoča **opis enot s porazdelitvami**.

Primer simbolnega opisa osebnega omrežja

Ego – izbrana oseba s svojim osebnim omrežjem (enota analize).
Opis ega: s spremenljivkami $U_k, k = 1, \ldots m_e$, merjenimi v različnih merskih lestvicah.

Alterji – z egom povezane osebe iz omrežja.
Opis alterja ali njegovega razmerja – povezave z egom: s (pogosto drugimi) spremenljivkami $V_l, l = 1, \ldots m_a$.



Opis ega

U_1 = zadovoljen z materialno oporo	U_2 = zakonski status	U_3 = koliko let že živi v mestu	
1 = zelo nezadovoljen	1 = samski	1 = 10 ali manj	6 = 31-35
2 = dokaj nezadovoljen	2 = živi kot poročen	2 = 11-18	7 = 36-40
3 = malo nezadovoljen	3 = poročen	3 = 19-21	8 = 41-47
4 = malo zadovoljen	4 = ločen	4 = 22-25	9 = 48-54
5 = dokaj zadovoljen	5 = ovdovel	5 = 26-30	10 = 55 ali več
6 = zelo zadovoljen			

Izbrani ego

$$\boldsymbol{X}^{o}{}_{1001} \equiv [\mathbf{u_1} = \text{ malo zadovoljen }, \mathbf{u_2} = \text{ samski }, \mathbf{u_3} = 24 \text{ let}]$$

ima simbolni opis

$$\boldsymbol{X}_{1001} = [\mathbf{4\ ,1\ ,4}] = [[0,0,0,1,0,0],[1,0,0,0,0],[0,0,0,1,0,0,0,0,0,0]].$$

Opisi alterjev

V_1 = **spol**	V_2 = **kako daleč od ega živi**
1 = moški	1 = 15 min ali manj
2 = ženski	2 = nad 15 do 30 min
	3 = nad 30 do 60 min
	4 = uro ali več

Z izbranim egom povezane osebe iz njegovega osebnega omrežja – alterji:

$$\begin{aligned} \boldsymbol{Y}^o{}_{1001:1} &\equiv [v_1 = \text{ženski}, v_2 = 25 \text{ min}] \\ \boldsymbol{Y}^o{}_{1001:2} &\equiv [v_1 = \text{ženski}, v_2 = 50 \text{ min}] \\ \boldsymbol{Y}^o{}_{1001:3} &\equiv [v_1 = \text{moški}, v_2 = 20 \text{ min}] \end{aligned}$$

Opisi alterjev:

$$\begin{aligned} \boldsymbol{Y}_{1001:1} &= [2,2] = \textbf{[[0,1], [0,1,0,0]]} \\ \boldsymbol{Y}_{1001:2} &= [2,3] = \textbf{[[0,1], [0,0,1,0]]} \\ \boldsymbol{Y}_{1001:3} &= [1,2] = \textbf{[[1,0], [0,1,0,0]]} \end{aligned}$$

Simbolni objekt ega X_{1001}

Osebno omrežje ega $\mathbf{X}_{1001}$:

$$\mathbf{A}(\mathbf{X}_{1001}) = \{\mathbf{Y}_{1001:1}, \mathbf{Y}_{1001:2}, \mathbf{Y}_{1001:3}\}$$

Opis omrežja s porazdelitvami:

$$So(\mathbf{A}(\mathbf{X}_{1001})) \quad = \quad [[\frac{\mathbf{1}}{\mathbf{3}}, \frac{\mathbf{2}}{\mathbf{3}}], [\mathbf{0}, \frac{\mathbf{2}}{\mathbf{3}}, \frac{\mathbf{1}}{\mathbf{3}}, \mathbf{0}]]$$

Simbolni objekt ega $\mathbf{X}_{1001}$:

$$So(\mathbf{X}_{1001}) \quad = \quad [\mathbf{4}\ \mathbf{,1}\ \mathbf{,4}, [\frac{\mathbf{1}}{\mathbf{3}}, \frac{\mathbf{2}}{\mathbf{3}}], [\mathbf{0}, \frac{\mathbf{2}}{\mathbf{3}}, \frac{\mathbf{1}}{\mathbf{3}}, \mathbf{0}]]$$

Različnost med enotama/skupinama

Utežena vsota različnosti po spremenljivkah

$$d(\mathbf{X}_1, \mathbf{X}_2) = \sum_{j=1}^{m} \alpha_j \, d(\mathbf{X}_1, \mathbf{X}_2; V^j), \qquad \sum_{j=1}^{m} \alpha_j = 1, \ \alpha_j \geq 0,$$

kjer je

$$d_{abs}(\mathbf{X}_1, \mathbf{X}_2; V^j) = \frac{1}{2} \sum_{i=1}^{k_j} |f(i, \mathbf{X}_1; V^j) - f(i, \mathbf{X}_2; V^j)|$$

ali

$$d_{sqr}(\mathbf{X}_1, \mathbf{X}_2; V^j) = \frac{1}{2} \sum_{i=1}^{k_j} (f(i, \mathbf{X}_1; V^j) - f(i, \mathbf{X}_2; V^j))^2.$$

Razvrščanje kot optimizacijski problem

Poišči dopustno razvrstitev $\mathcal{C}^*$, pri kateri je

$$P(\mathcal{C}^*) = \min_{\mathcal{C} \in \Phi} P(\mathcal{C}).$$

Množica dopustnih razvrstitev je **množica razbitij** končne množice enot $\mathbf{E}$.

Kriterijska funkcija je vsota napak skupin

$$P(\mathcal{C}) = \sum_{C \in \mathcal{C}} p(C),$$

kjer je

$$p(C) = \sum_{\mathbf{X} \in C} d(\mathbf{X}, \mathbf{L}_C).$$

$\mathbf{L}_C$ je **voditelj** ali **predstavnik** skupine C.

Prilagojena metoda voditeljev

Osnovna shema:

izberi začetno razvrstitev

repeat

vsaki skupini $C \in \mathcal{C}$ določi voditelja $\mathbf{L}_C$;

pridruži vsako enoto najbližjemu voditelju

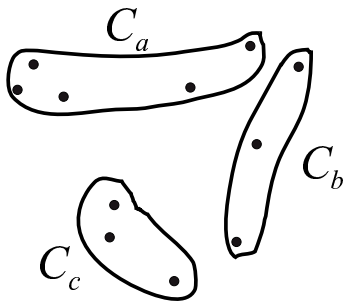
until voditelji se ustalijo.

Opis voditelja

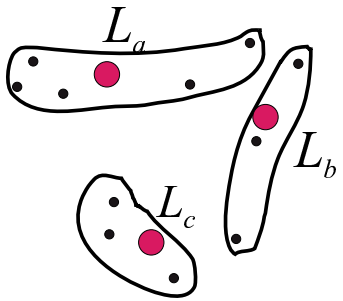
$$\boldsymbol{L} = [\boldsymbol{L}(V^1), \ldots, \boldsymbol{L}(V^m)],$$

$$\boldsymbol{L}(V^j) = [s(1, \mathbf{L}; V^j), \ldots, s(k_j, \mathbf{L}; V^j)], \quad \sum_{i=1}^{k_j} s(i, \mathbf{L}; V^j) = 1.$$

Določitev novih voditeljev

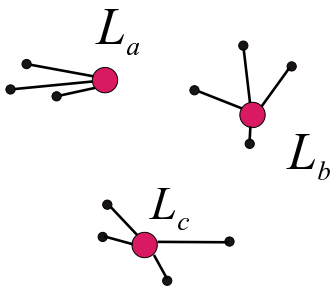


Stare skupine enot

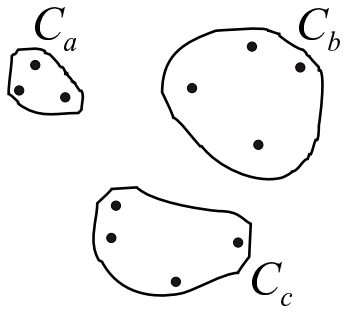


Novi voditelji skupin

Določitev novih skupin



Stari voditelji skupin



Nove skupine

Optimalni voditelj za P_{abs}

IZREK 1 *(izvirni prispevek)*

Kadar ***vse enote pri vsaki spremenljivki zavzamejo le posamezno vrednost****, je optimalni voditelj skupine za kriterijsko funkcijo* P_{abs} *z različnostjo* d_{abs} *določen z* ***največjimi frekvencami***

$$s(i, \mathbf{L}_C; V) = \begin{cases} \frac{1}{t} & \textit{za } \ i \in M \\ 0 & \textit{sicer} \end{cases}$$

kjer je t *število vseh razredov z največjimi frekvencami pri spremenljivki* V *in* $M = \{j : q(j, C; V) = \max\limits_{1 \le l \le k_V} q(l, C; V)\}$.

Optimalni voditelj za P_{sqr}

IZREK 2 *(izvirni prispevek)*

Za kriterijsko funkcijo P_{sqr} z različnostjo d_{sqr} je optimalni voditelj skupine določen s ***povprečjem relativnih frekvenc***

$$s(i, \mathbf{L}_C; V) = \frac{1}{\mathrm{card}(C)} \sum_{\mathbf{X} \in C} f(i, \mathbf{X}; V).$$

Primer opisa optimalnega voditelja za P_{abs}

Podatki o avtomobilih

Skupina C:

$q(C; cena)$	0	0	0	1	5	42		
$q(C; tip)$	0	0	0	0	2	23	0	23
$q(C; st.vrat)$	46	2	0	0				

Optimalni voditelj pri kriterijski funkciji P_{abs}:

$s(\mathbf{L}_C; cena)$	0	0	0	0	0	1		
$s(\mathbf{L}_C; tip)$	0	0	0	0	0	$\frac{1}{2}$	0	$\frac{1}{2}$
$s(\mathbf{L}_C; st.vrat)$	1	0	0	0				

Primer opisa optimalnega voditelja za P_{sqr}

Optimalni voditelj pri kriterijski funkciji P_{sqr}:

$$
\begin{array}{lcccccccc}
s(\mathbf{L}_C; cena) & 0 & 0 & 0 & \frac{1}{48} & \frac{5}{48} & \frac{42}{48} & & \\
s(\mathbf{L}_C; tip) & 0 & 0 & 0 & 0 & \frac{2}{48} & \frac{23}{48} & 0 & \frac{23}{48} \\
s(\mathbf{L}_C; st.vrat) & \frac{46}{48} & \frac{2}{48} & 0 & 0 & & & &
\end{array}
$$

Značilnosti skupine C:

cena	=	od 6 550 000 do 45 000 000 SIT	87.50 %
tip	=	RO (športni dvosed z zložljivo streho) ali	47.92 %
		KU (zaprt dvosedežen avto)	47.92 %
št.vrat	=	2	95.83 %

Prednosti izbire P_{abs}

- **hitrejša konvergenca**,
- **enostavnejše tolmačenje** predstavnika in značilnosti skupine.

Pomanjkljivosti izbire P_{abs}

- enote morajo biti pri vseh spremenljivkah podane le s posameznimi vrednostmi,
- v splošnem so dobljene skupine manj enovite kot pri izbiri P_{sqr}.

Prednosti izbire P_{sqr}

- vhodne **enote so lahko podane s porazdelitvami**, in ne le z eno samo vrednostjo,
- **optimalni voditelj je enolično določen**,
- kadar so **enote pri vseh spremenljivkah podane s posameznimi vrednostmi**, je optimalni voditelj skupine opisan z **relativnimi frekvencami skupine**

$$s(i, \mathbf{L}_C; V) = f(i, C; V).$$

Prilagojena hierarhična metoda združevanja

Osnovna shema:

vsaka enota je skupina: $\mathcal{C}_n = \{\{\mathbf{X}\}: \mathbf{X} \in \mathbf{E}\}$;
nivo začetnih skupin je 0: $h(\{\mathbf{X}\}) = 0,\ \mathbf{X} \in \mathbf{E}$;
for $k := n$ **downto** 2 **do**
poišči najbližji skupini in ju združi:
$(p, q) = \underset{i,j:i\neq j}{\operatorname{argmin}}\{D(C_i, C_j): C_i, C_j \in \mathcal{C}_k\}$;
$\mathcal{C}_{k-1} = \mathcal{C}_k \setminus \{C_p, C_q\} \cup \{C_p \cup C_q\}$;
določi nove različnosti $D(C_p \cup C_q, C)$ za vse $C \in \mathcal{C}_k \setminus \{C_p, C_q\}$;
nivo nove skupine: $h(C_p \cup C_q) = D(C_p, C_q)$;
endfor

Hierarhične metode kot požrešne metode

Definicija (Batagelj, Ferligoj)

Kriterijska funkcija P je **usklajena** (angl. compatible) z različnostjo med skupinami D natanko tedaj, ko velja:

$$\begin{aligned} P(\mathcal{C}) &= \bigoplus_{C \in \mathcal{C}} p(C) \\ p(\{\mathbf{X}\}) &= 0,\ \forall \mathbf{X} \in \mathbf{E} \\ p(C) &= \min_{C' : \emptyset \subset C' \subset C} (p(C') \oplus p(C \setminus C') \oplus D(C', C \setminus C')), \end{aligned}$$

kjer je $(\mathbb{R}_0^+, \bigoplus, 0, \leq)$ urejen Abelov monoid.

Primeri usklajenih kriterijskih funkcij in mer različnosti

METODA	$\oplus$	$p(C)$	$D(C_1, C_2)$
maksimalna (CL)	max	$\max_{\mathbf{X},\mathbf{Y}\in C} d(\mathbf{X},\mathbf{Y})$	$\max_{\mathbf{X}\in C_1,\mathbf{Y}\in C_2} d(\mathbf{X},\mathbf{Y})$
minimalna (SL)	+	vrednost minimalnega vpetega drevesa nad C z vrednostmi povezav d	$\min_{\mathbf{X}\in C_1,\mathbf{Y}\in C_2} d(\mathbf{X},\mathbf{Y})$
Wardova	+	$\sum_{X\in C} \|\|X-\overline{C}\|\|^2$	$\frac{n_1\cdot n_2}{n_1+n_2}\|\|\overline{C_1}-\overline{C_2}\|\|^2$

$$n_i = card(C_i), \overline{C_i} = \frac{1}{n_i}\sum_{X\in C_i} X$$

Wardova metoda predpostavlja vektorski opis enot s številskimi komponentami

Osnovni izrek

IZREK 3 *(Batagelj)*

Za kriterijsko funkcijo P, ki je usklajena z različnostjo med skupinami D, velja:

$$P(\mathcal{C}_k^*) = \min_{\mathcal{C} \in \Phi_{k+1}} (P(\mathcal{C}) \oplus \min_{C_i, C_j \in \mathcal{C}} D(C_i, C_j)).$$

S $\mathcal{C}_k^*$ smo označili optimalno razbitje množice enot na k skupin glede na vrednost kriterijske funkcije P (v razbitju $\mathcal{C}_k^*$ je vrednost kriterijske funkcije P na Φ_k najmanjša).

Požrešna 'aproksimacija'

Na osnovi zveze

$$P(\mathcal{C}_k^*) = \min_{\mathcal{C} \in \Phi_{k+1}} (P(\mathcal{C}) \oplus \min_{C_i, C_j \in \mathcal{C}} D(C_i, C_j))$$

izpeljemo požrešno različico hierarhične metode združevanja:

$$P(\mathcal{C}_k^\bullet) = P(\mathcal{C}_{k+1}^\bullet) \oplus D(C_p, C_q),$$

kjer je

$$D(C_p, C_q) = \min_{C_i, C_j \in \mathcal{C}_{k+1}^\bullet} D(C_i, C_j)$$

in je $\mathcal{C}_{k+1}^\bullet$ rešitev, dobljena po požrešni metodi na prejšnjem koraku združevanja.

Nivo skupine

Grafično predstavitev hierarhične razvrstitve imenujemo **dendrogram**.

Nivo $h(C)$ skupine $C = C_p \cup C_q$ je določen z različnostjo med skupinama C_p and C_q

$$h(C_p \cup C_q) = D(C_p, C_q).$$

$h(C) = 0$ kadar je C voditelj ali začetna skupina.

Velja zveza

$$D(C_p, C_q) = P(\mathcal{C}_k^\bullet) - P(\mathcal{C}_{k+1}^\bullet) = p(C_p \cup C_q) - p(C_p) - p(C_q).$$

Nivo skupine za P_{sqr}

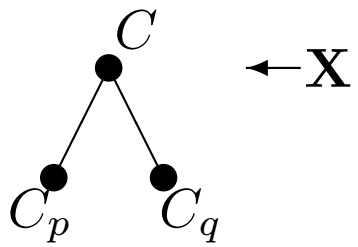
IZREK 4 *(izvirni prispevek)*

Pri drugi kriterijski funkciji ***P_{sqr}*** *z različnostjo* d_{sqr} *se računanje različnosti med skupinama* $D(C_p, C_q)$ *poenostavi v znano* ***Wardovo relacijo***

$$D_{sqr}(C_p, C_q) = \frac{\mathrm{card}(C_p) \cdot \mathrm{card}(C_q)}{\mathrm{card}(C_p) + \mathrm{card}(C_q)} \, d_{sqr}(\mathbf{L}_p, \mathbf{L}_q).$$

Metoda dodajanja

Enoto $\mathbf{X}$ vstavljamo v tekoče vozlišče $C = C_p \cup C_q$:



Za enoto $\mathbf{X}$ obstajajo tri možnosti, kam jo lahko spustimo v drevesu:

1. enoto $\mathbf{X}$ pridružimo skupini C_p,
2. enoto $\mathbf{X}$ pridružimo skupini C_q,
3. enota $\mathbf{X}$ ustvari novo vozlišče (izbirno – če dovolimo dodajanje novih vozlišč v drevo).

Maksimalna različnost med skupinama

Iz požrešne različice

$$P(\mathcal{C}_k^\bullet) = P(\mathcal{C}_{k+1}^\bullet) + D(C_p, C_q)$$

izpeljemo relacijo za hierarhično metodo dodajanja

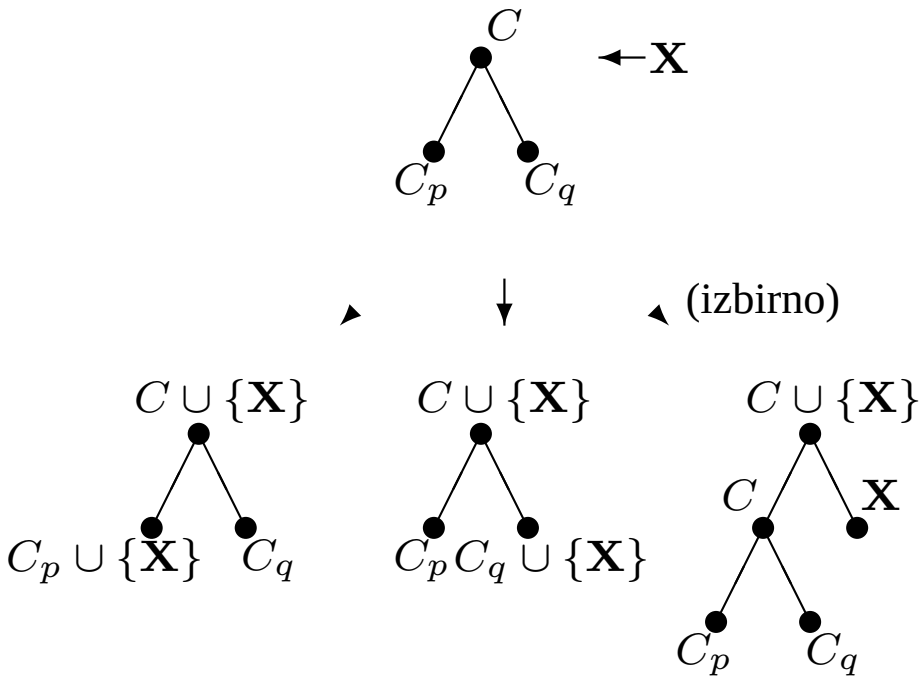
$$P(\mathcal{C}_{k+1}^\bullet) = P(\mathcal{C}_k^\bullet) - D(C_p, C_q).$$

$\mathcal{C}_k^\bullet$ je rešitev, dobljena po požrešni metodi na prejšnjem koraku dodajanja. Vrednosti kriterijske funkcije $P(\mathcal{C}_{k+1}^\bullet)$ bo najmanjša pri izbiri skupin C_p, C_q, kjer je $C_p \cup C_q \in \mathcal{C}_k^\bullet$ in

$$D(C_p, C_q) = \max_{C_i, C_j : C_i \cup C_j \in \mathcal{C}_k^\bullet} D(C_i, C_j).$$

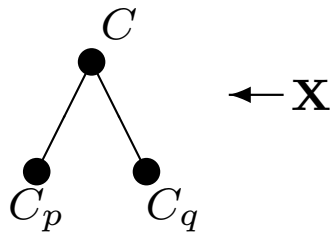
Kam naprej? Požrešnost

IZBIRA $\max\{D(C_p \cup \{\mathbf{X}\}, C_q), D(C_p, C_q \cup \{\mathbf{X}\}), D(C, \{\mathbf{X}\})\}$

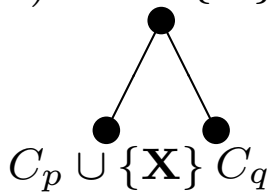


Kam naprej? Običajni pristop

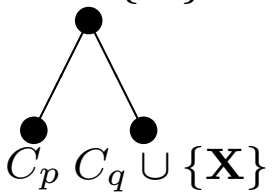
IZBIRA najbližjega sina: $\min\{D(C_p, \{\mathbf{X}\}), D(C_q, \{\mathbf{X}\})\}$



Če je $D(C_p, \mathbf{X}) < D(C_q, \mathbf{X})$: $C \cup \{\mathbf{X}\}$



Sicer: $C \cup \{\mathbf{X}\}$



Primerjava obeh načinov spuščanja enote po drevesu

IZREK 5

Pri enakem začetnem drevesu in izključitvi možnosti dodajanja novih vozlišč v drevo sta načina spuščanja enote po drevesu (izbira največje različnosti med skupinama ali izbira najbližjega sina) enakovredna.

Zastave

Vir: *Collins Gem Guide to Flags*: Collins Publishers (1986).

Podatke prispeval: Richard S. Forsyth. 15. 5. 1990

Spletna naslova slikovnih prikazov: *Flags of The World*

`http://fotw.digibel.be/flags`,
`http://www.fotw.ca/flags/`

Število enot: **194**, število spremenljivk: **23**.

Spremenljivke:

NAVPIČNIC (angl. bars): število navpičnic (vertikal),

VODORAVNIC (angl. stripes): število vodoravnic (horizontal),

ŠTEVILO BARV (angl. colours): število različnih barv na zastavi,

RDEČA, ZELENA, MODRA, ZLATA ali RUMENA, BELA, ČRNA,

ORANŽNA ali RJAVA: 0 – ni prisotna, 1 – je prisotna,

BARVA OSNOVE (angl. mainhue): osnovna barva zastave,

KROŽNIC (angl. circles): število krožnic na zastavi,

KRIŽEV (angl. crosses): število pokončnih križev,

ANDREJEVIH KRIŽEV (angl. saltires): število diagonalnih križev,

ČETRTIN (angl. quarters): število razčetverjenih delov,

SONC ali ZVEZD (angl. sunstars): število sonc ali zvezd,

POLMESEC (angl. crescent): 0 – luna oz. mesec ni prisoten, 1 – je prisoten,

TRIKOTNIK: 0 – ni prisoten, 1 – je prisoten,

IKONA (predstavlja neživo stvar ali njen del (npr. čoln)): 0 – ni prisotna, 1 – je prisotna,

ANIMIRANI SIMBOL (angl. animate): 0 – animirana ikona (npr. orel, drevo, človeška roka) ni prisotna, 1 – je prisotna,

BESEDILO: 1 – vsaj črka je prisotna, 0 – ni prisotna,

ZG.LEVO (angl. topleft): barva na zgornjem levem robu,

SP.DESNO (angl. botright): barva na spodnjem desnem robu.

Zastave 1. skupine

Modusne vrednosti 1. skupine zastav

spr.	modus	frekv.	rel.frekv. v %	spr.	modus	frekv.	rel.frekv. v %
rdeča	da	29	100,00 %	trikotnik	ne	25	86,21 %
polmesec	ne	29	100,00 %	navpičnic	0	25	86,21 %
križev	0	29	100,00 %	zelena	da	23	79,31 %
Andr. križev	0	29	100,00 %	ikona	ne	23	79,31 %
oranžna	ne	28	96,55 %	zg. levo	rdeča	20	68,97 %
četrtink	0	28	96,55 %	črna	da	20	68,97 %
krožnic	0	28	96,55 %				
modra	ne	26	89,66 %				
besedilo	ne	26	89,66 %				
anim. simbol	ne	26	89,66 %				

Zastave 2. skupine

Modusne vrednosti 2. skupine zastav

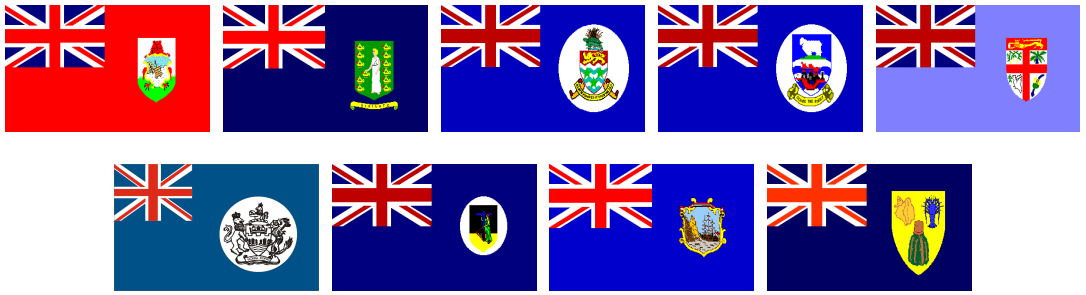
spr.	modus	frekv.	rel.frekv. v %	spr.	modus	frekv.	rel.frekv. v %
križev	0	33	100, 00 %	oranžna	ne	26	78, 79 %
Andr. križev	0	33	100, 00 %	navp.	0	26	78, 79 %
rdeča	da	32	96, 97 %	modra	da	25	75, 76 %
četrtink	0	32	96, 97 %	krožnic	0	24	72, 73 %
polmesec	ne	32	96, 97 %	sp. desno	rdeča	24	72, 73 %
zlata (rum.)	da	30	90, 91 %	bela	da	23	69, 70 %
trikotnik	ne	27	81, 82 %	zelena	ne	23	66, 67 %
besedilo	ne	27	81, 82 %				

Zastave 3. skupine

Modusne vrednosti 3. skupine zastav

spr.	modus	frekv.	rel.frekv. v %	spr.	modus	frekv.	rel.frekv. v %
modra	da	47	100,00 %	trikotnik	ne	43	91,49 %
besedilo	ne	47	100,00 %	krožnic	0	43	91,49 %
zelena	ne	46	97,87 %	Andr. križev	0	40	85,11 %
ikona	ne	46	97,87 %	zlata (rum.)	ne	38	80,85 %
anim.simb.	ne	46	97,87 %	križev	0	34	72,34 %
bela	da	45	95,74 %	četrtink	0	33	70,21 %
polmesec	ne	45	95,74 %	sp. desno	modra	31	65,96 %
oranžna	ne	44	93,62 %				
črna	ne	44	93,62 %				
navpičnic	0	44	93,62 %				

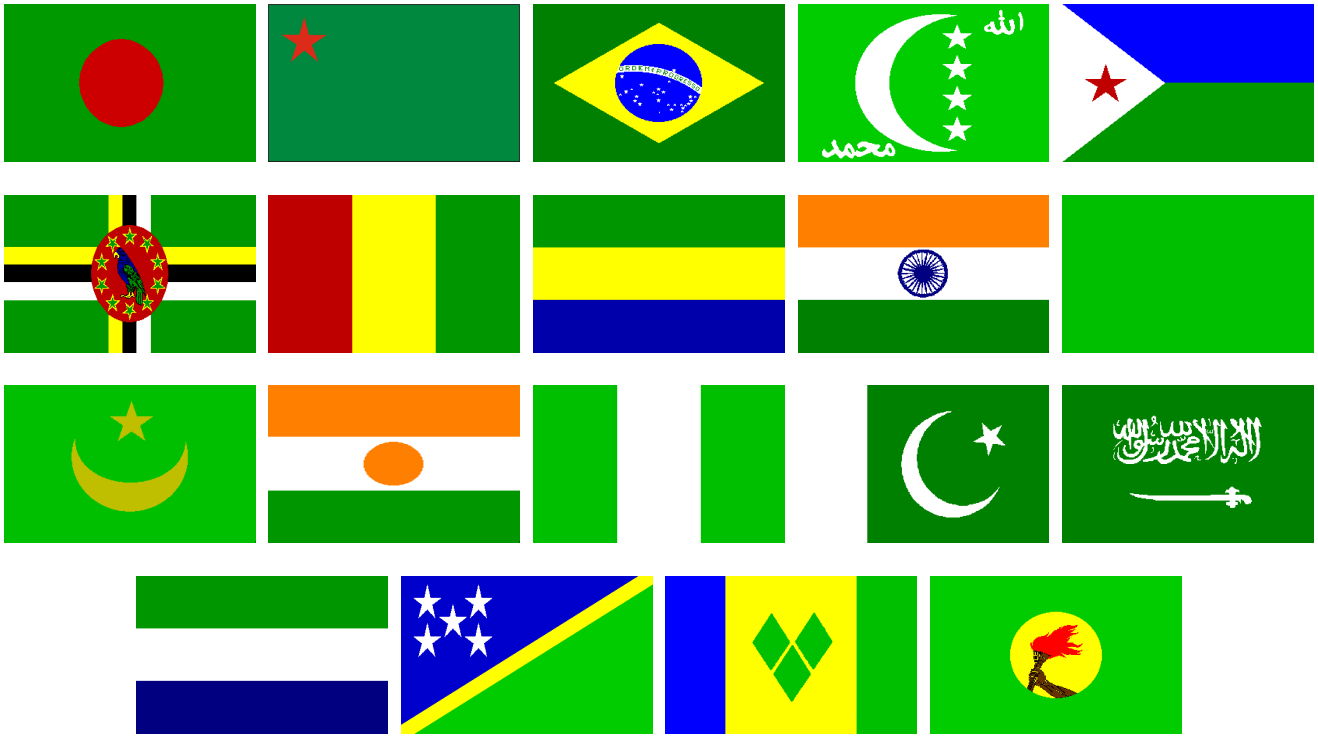
Zastave 4. skupine



Modusne vrednosti 4. skupine zastav

spr.	modus	frekv.	rel.frekv. v %	spr.	modus	frekv.	rel.frekv. v %
modra	da	9	100, 00 %	četrtink	1	9	100, 00 %
rdeča	da	9	100, 00 %	ikona	da	9	100, 00 %
bela	da	9	100, 00 %	navp.	0	9	100, 00 %
zlata (rum.)	da	9	100, 00 %	vodorav.	0	9	100, 00 %
zelena	da	9	100, 00 %	trikotnik	ne	9	100, 00 %
Andr. križev	1	9	100, 00 %	polmesec	ne	9	100, 00 %
zg. levo	bela	9	100, 00 %				

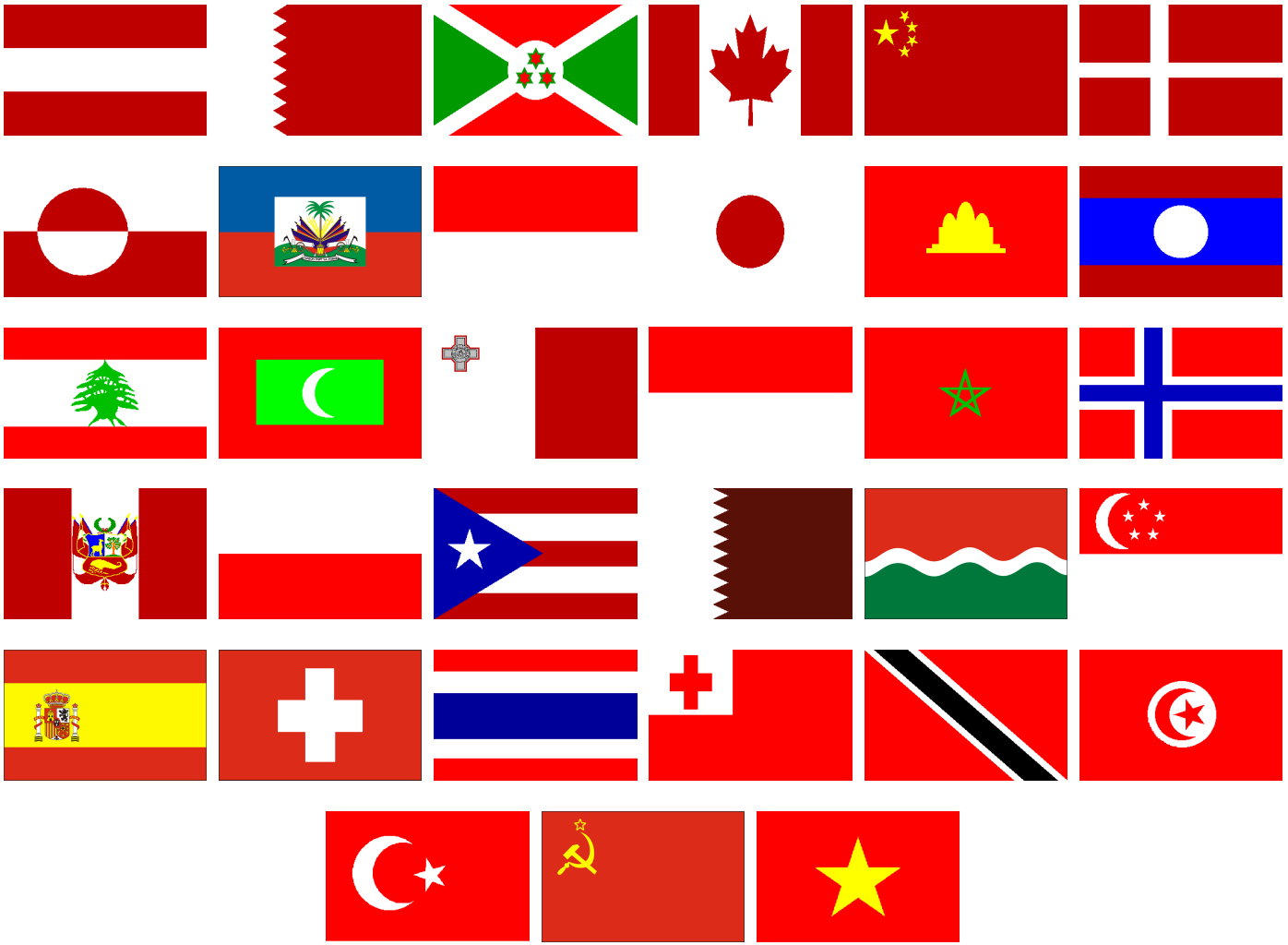
Zastave 5. skupine



Modusne vrednosti 5. skupine zastav

spr.	modus	frekv.	rel.frekv. v %	spr.	modus	frekv.	rel.frekv. v %
zelena	da	19	100, 00 %	oranžna	ne	16	84, 21 %
Andr. križev	0	19	100, 00 %	navp.	0	16	84, 21 %
četrtink	0	19	100, 00 %	besedilo	ne	16	84, 21 %
križev	0	19	100, 00 %	ikona	da	15	78, 95 %
črna	ne	18	94, 74 %	vodorav.	0	14	73, 68 %
osnova	zelena	16	84, 21 %	bela	da	13	68, 42 %
sp. desno	zelena	16	84, 21 %	zg. levo	zelena	13	68, 42 %
anim. simb.	ne	16	84, 21 %	rdeča	ne	13	68, 42 %
trikotnik	ne	16	84, 21 %	krožnic	ne	13	68, 42 %
polmesec	ne	16	84, 21 %				

Zastave 6. skupine



Modusne vrednosti 6. skupine zastav

spr.	modus	frekv.	rel.frekv. v %	spr.	modus	frekv.	rel.frekv. v %
besedilo	ne	33	100, 00 %	zelena	ne	28	84, 85 %
rdeča	da	32	96, 97 %	zlata (rum.)	ne	28	84, 85 %
Andr. križev	0	32	96, 97 %	krožnic	0	28	84, 85 %
četrtink	0	32	96, 97 %	križev	0	28	84, 85 %
trikotnik	ne	31	93, 94 %	osnova	rdeča	27	81, 82 %
oranžna	ne	31	93, 94 %	bela	da	26	78, 79 %
anim. simb.	ne	31	93, 94 %	sp. desno	rdeča	25	75, 76 %
ikona	ne	30	90, 91 %	zg. levo	rdeča	23	69, 70 %
črna	ne	30	90, 91 %	št. barv	2	23	69, 70 %
polmesec	ne	29	87, 88 %	sonc, zv.	0	23	69, 70 %
navpičnic	0	29	87, 88 %	vodorav.	0	23	69, 70 %
modra	ne	29	87, 88 %				

Zastave 7. skupine

Modusne vrednosti 7. skupine zastav

spr.	modus	frekv.	rel.frekv. v %	spr.	modus	frekv.	rel.frekv. v %
besedilo	ne	16	100,00 %	črna	ne	14	87,50 %
trikotnik	ne	16	100,00 %	vodorav.	0	14	87,50 %
Andr. kr.	0	16	100,00 %	rdeča	da	13	81,25 %
četrtink	0	16	100,00 %	ikona	ne	13	81,25 %
krožnic	0	16	100,00 %	sonc, zv.	0	13	81,25 %
križev	0	16	100,00 %	št. barv	3	12	75,00 %
zelena	da	15	93,75 %	oranžna	ne	12	75,00 %
modra	ne	15	93,75 %	anim. simb.	ne	12	75,00 %
polmesec	ne	15	93,75 %				

Zastave 8. skupine

Modusne vrednosti 8. skupine zastav

spr.	modus	frekv.	rel.frekv. v %	spr.	modus	frekv.	rel.frekv. v %
zelena	da	8	100, 00 %	Andr. križev	0	7	87, 50 %
zlata (rum.)	da	8	100, 00 %	modra	ne	7	87, 50 %
črna	da	8	100, 00 %	osnova	zelena	6	75, 00 %
trikotnik	da	8	100, 00 %	rdeča	da	6	75, 00 %
oranžna	ne	8	100, 00 %	ikona	ne	6	75, 00 %
besedilo	ne	8	100, 00 %	anim. simb.	ne	6	75, 00 %
polmesec	ne	8	100, 00 %				
navp.	0	8	100, 00 %				
četrtink	0	8	100, 00 %				
krožnic	0	8	100, 00 %				
križev	0	8	100, 00 %				

Mednarodni sociološki podatki raziskave ISSP 1988, 1994

Vir: *The International Social Survey Programme ISSP 1985 – 1995, Data and Documentation*. Special edition for the international conference on Large Scale Data Analysis at the Zentralarchiv, Cologne, May 25-28, 1999.

Število enot: **45 784**, število spremenljivk: **34**.

Tema: *DRUŽINA IN VLOGA OČETOV IN MATER V DRUŽINI*
Family and Changing Gender Roles I - 1988 (ZA-No. 1700): 8 držav, skupno 12 194 enot
Family and Changing Gender Roles II - 1994 (ZA-No. 2620): 24 držav, skupno 33 590 enot

1. del analize: razvrščanje celotnega podatkovja (metoda voditeljev)

Primer opisa značilnosti ene od skupin, skupine C_{19}:

% enot	spremenljivka	najpogostejša vrednost
$80,13\,\%$	Oba starša morata prispevati v gospod. prorač. (V11)	se strinjam
$75,60\,\%$	Otroci: odraščanje veselje (V29)	se strinjam
$74,25\,\%$	Vprašanec kdaj ločen? (V38)	ne
$71,84\,\%$	Delo pripomore k neodvisnosti žensk (V9)	se strinjam
$70,14\,\%$	Poroka: bolje če so otroci zaželjeni (V23)	se strinjam
$69,97\,\%$	Ženske resnično želijo dom in otroke (V7)	se strinjam
$69,62\,\%$	Zakonski stan (V202)	poročen/živi kot poročen
$68,15\,\%$	Naj se ženska zaposli: otroke pusti doma (V18)	dela polni delovni čas
$67,56\,\%$	Zaposlena mati:prikrajšana družina (V6)	se strinjam
$67,35\,\%$	Zaposlena mati:predšolski otrok prikrajšan (V5)	se strinjam
$66,71\,\%$	Soprog/soproga kdaj ločen/a? (V39)	ne

Deleži držav

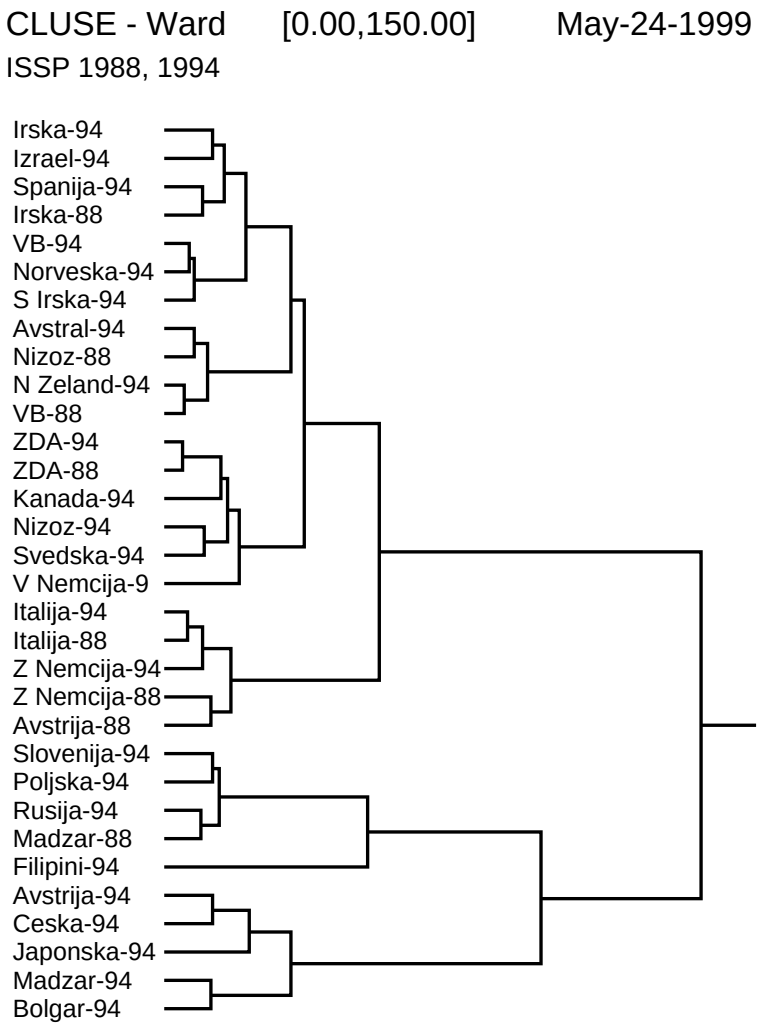
Država	št. enot	% v državi	% v skupini
Avstralija-94	96	5,396	1,802
Avstrija-88	121	12,449	4,158
Avstrija-94	30	3,071	1,026
Bolgarija-94	23	2,043	0,682
Češka-94	33	3,223	1,076
Filipini-94	656	54,667	**18,257**
Irska-88	124	12,338	4,121
Irska-94	125	13,326	4,451
Italija-88	65	6,323	2,112
Italija-94	66	6,483	2,165
Izrael-94	151	11,733	3,918
Japonska-94	37	2,831	0,945
Kanada-94	34	2,361	0,789
Madžarska-88	425	24,467	**8,172**
Madžarska-94	60	4,000	**1,336**
Nizozemska-88	63	3,627	1,211
Nizozemska-94	93	4,726	1,578

Deleži držav...

Država	št. enot	% v državi	% v skupini
Norveška-94	113	5,414	1,808
Nova Zelandija-94	47	4,489	1,499
Poljska-94	301	18,848	**6,295**
Rusija-94	395	19,770	**6,603**
Severna Irska-94	56	8,655	2,891
Slovenija-94	202	19,574	**6,537**
Španija-94	331	13,272	4,432
Švedska-94	43	3,381	1,129
Velika Britanija-88	65	4,973	1,661
Velika Britanija-94	57	5,793	1,935
Vzhodna Nemčija-94	27	2,461	0,822
ZDA-88	67	4,738	1,582
ZDA-94	77	5,321	1,777
Zahodna Nemčija-88	113	3,774	1,261
Zahodna Nemčija-94	137	5,895	1,969

2. del analize:

Hierarhija nad državami glede na zastopanost držav v 22. skupinah



Osebna (ego-centrična) omrežja

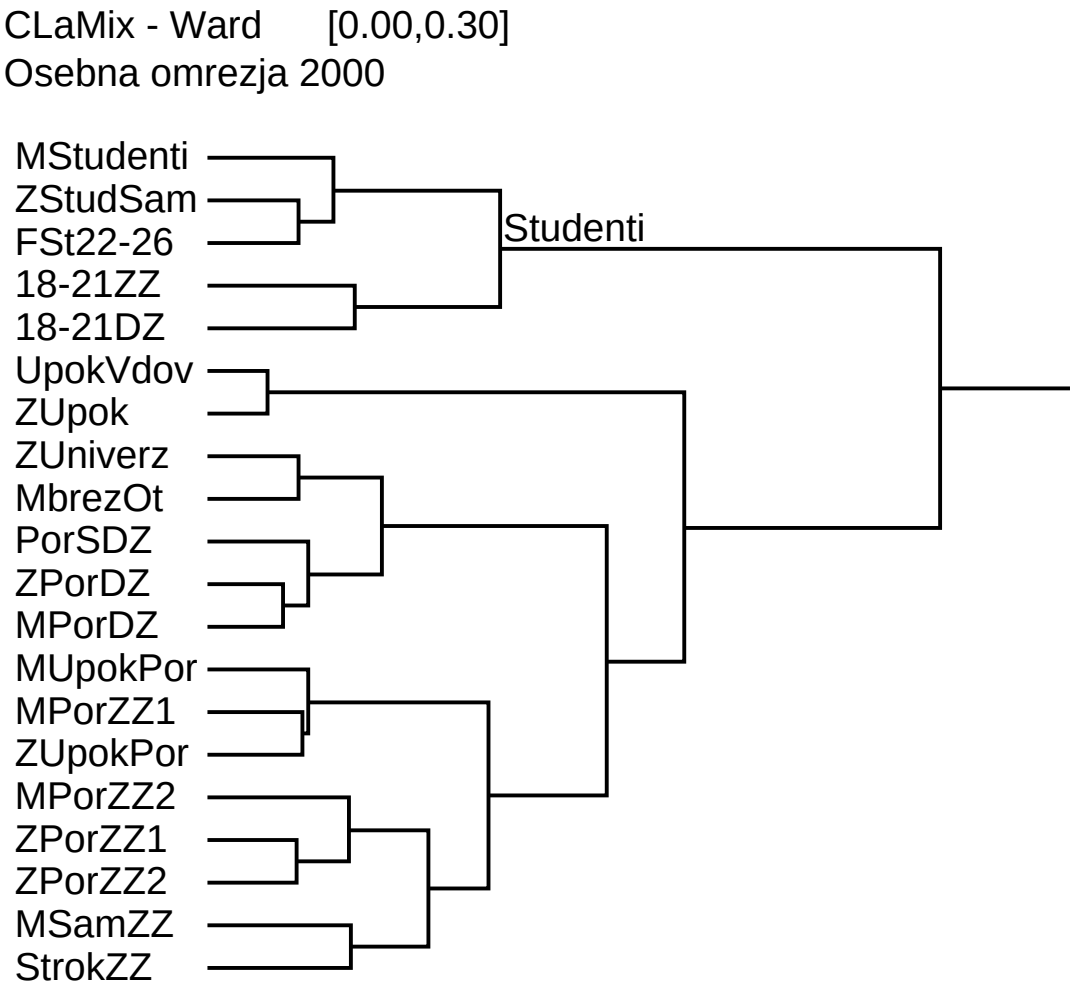
Vir: Omrežje socialne opore, zbrano na Fakulteti za družbene vede v Ljubljani med marcem in junijem leta 2000 s telefonskimi intervjuji (CATI).

Število enot: **1032 (+ 5849)**, število spremenljivk: **38 (+ 10)**.

Zastavljeni cilji:

1. smiselno *zmanjšati veliko podatkovje*,
2. *vključiti v proces vse spremenljivke* (merjene v različnih merskih lestvicah),
3. *ohraniti čim več informacij*.

Dendrogram nad dobljenimi skupinami



Opis skupine Studenti

Porazdelitve za prve štiri spremenljivke Q1a, Q2a, Q3a in Q4a:

V	zelo nez.	dokaj nez.	nekoliko nez.	nekoliko zad.	dokaj zad.	zelo zad.	manjka
Q1a:	0	0	4	8	79	118	6
Q2a:	0	0	0	4	71	135	5
Q3a:	0	0	0	9	73	131	2
Q4a:	0	0	0	7	68	134	6

Značilnosti skupine Studenti

Demografske spremenljivke:

- $85,58\ \%$ D10(status) = samski
- $74,42\ \%$ D9(izobrazba) = srednja šola
- $63,26\ \%$ D8(poklic) = študent
- $59,53\ \%$ D2(koliko otrok mlajših od 19 let v gospodinjstvu) = nič
- $52,56\ \%$ SPOL = ženski

Spremenljivke, s katerimi ego izraža zadovoljstvo nad vrstami opore alterjev:

- $62,79\ \%$ Q4a(čustvena opora) = zelo zadovoljen
- $62,33\ \%$ Q2a(informacijska opora) = zelo zadovoljen
- $60,93\ \%$ Q3a(socialna opora) = zelo zadovoljen
- $54,88\ \%$ Q1a(materialna opora) = zelo zadovoljen

Spremenljivka alterjev:

- $50,71\ \%$ Q12(spol alterjev) = moški

Zaključek

Lastnosti v disertaciji razvitih metod:

- primerne so za razvrščanje **velikih** podatkovij z **neenovitimi** opisi enot;
- enote so lahko **opisane s porazdelitvami**, ne le z vektorji posameznih vrednosti;
- omogočajo **utežitev spremenljivk** po pomembnosti;
- porazdelitve ohranijo **bogatejšo informacijo** o skupini;
- **preprosta razlaga** rezultatov zaradi ohranjene porazdelitve in dokaj enostavnega opisa optimalnega voditelja;
- **usklajenost predstavitev enote, skupine in voditelja** in **usklajenost računanja različnosti** s kriterijsko funkcijo omogočajo kombinirano uporabo razvitih metod.

Prispevki k znanosti

1. **Uporaba enovite predstavitve mešanih, neenovito opisanih podatkov** s simbolnimi objekti **v procesu razvrščanja** podatkov v skupine. Enovita prestavitev temelji na porazdelitvah po razredih zaloge vrednosti vsake od spremenljivk, s katerimi so podatki opisani.
2. **Teoretična izpeljava optimalnega predstavnika skupine** pri zastavitvi problema razvrščanja nad enovitim opisom enot (primerov) kot optimizacijskega problema. Optimalni predstavnik je določen bodisi z maksimalnimi frekvencami (pri različnosti d_{abs}) bodisi s povprečjem relativnih frekvenc (pri različnosti d_{sqr}).
3. **Razvoj učinkovitega algoritma** za razvrščanje **velike množice** podatkov z linearno časovno zahtevnostjo glede na produkt števila podatkov in števila skupin. Implementacija razvitih algoritmov v sistemu CLaMix.
4. **Prilagoditev hierarhične metode združevanja** na množice podatkov, ki so opisani s porazdelitvami, in dokaz, da je prilagoditev pri posebni izbiri različnosti posplošitev Wardove metode.
5. **Prilagoditev hierarhične metode dodajanja**.
6. Prikaz **povezane uporabe razvitih metod** (npr. zmanjšanje velikosti podatkovja z metodo voditeljev, prikaz notranje strukture skupin z dendrogramom).

Objave in predstavitve na konferencah

1. Zbornik z recenziranimi razširjenimi članki:
   - IFCS 1998 v Rimu, Italija: *Clustering large datasets of mixed units* (soavt. s prof. dr. V. Batageljem)
   - IFCS 2002 v Krakovu, Poljska: *Symbolic data analysis approach to clustering large datasets* (soavt. s prof. dr. V. Batageljem)
2. V obliki člankov so opisane metode razvrščanja predstavljene tudi v zbornikih posvetovanj DSI 1999 in 2001, v zbirki Metodološki zvezki 17, Developments in Statistics, 2002, posebni izdaji revije Informatica, Special Issue: Information Society and Inteligent Systems, 1999, zborniku 5th International Conference on Logic and Methodology, Cologne, Germany, 2000, zborniku Srečanja mladih statistikov v Vidmu YSMU 2000 in zborniku 4th European Conference on Principles and Practice of Knowledge Discovery in Databases.
3. Začetno delo avtorice disertacije na področju razvrščanja je objavljeno v soavtorstvu s prof. dr. Batageljem in prof. dr. Klavžarjem v članku *Dynamic Programming and Convex Clustering*, objavljenem v reviji Algorithmica, 1994.
4. Z objavljenimi povzetki so bile metode predstavljene še na več drugih mednarodnih konferencah.