Grupiranje podataka: klasteri

Kristian Sabo* Rudolf Scitovski† Ivan Vazler‡

Sažetak. *U ovom radu razmatramo problem grupiranja elemenata skupa $\mathcal{A}$ u disjunktne neprazne podskupove - klastere, pri čemu pretpostavljamo da su elementi skupa $\mathcal{A}$ određeni s jednim ili dva obilježja. Za rješavanje problema koristi se kriterij najmanjih kvadrata te kriterij najmanjih apsolutnih udaljenosti. Naveden je niz primjera koji ilustriraju razlike među tim kriterijima. Izrađena je odgovarajuća programska podrška s ciljem da zainteresirani stručnjaci u svom znanstvenom ili stručnom radu mogu olakšano koristiti ovu metodologiju i pristup.*

Ključne riječi: *grupiranje podataka-klasteri, aritmetička sredina, medijan, optimizacija*

Data clustering

Abstract. *In this paper we consider a clustering problem for a data-points set $\mathcal{A}$ into disjoint nonempty subsets - clusters, whereby it is assumed that elements of the set $\mathcal{A}$ are determined by one or two characteristics. Least square criteria and least absolute deviation criteria are used for solving the problem. A number of examples illustrating differences between these criteria are given. Corresponding software support is developed for the purpose of facilitating scientific or professional work by using this methodology and approach.*

Key words: *clusters, arithmetic mean, median, optimization*

1. Uvod

U ovom radu razmatramo problem grupiranja elemenata nekog skupa $\mathcal{A}$ s $m \geq 2$ elemenata u disjunktne podskupove $\pi_1, \ldots, \pi_k$, $1 \leq k \leq m$, takve da vrijedi

$$\bigcup_{i=1}^{k} \pi_i = \mathcal{A}, \qquad \pi_i \bigcap \pi_j = \emptyset, \quad i \neq j, \qquad m_j := |\pi_j| \geq 1, \quad j = 1, \ldots, k, \tag{1}$$

na osnovi jednog ili više obilježja uz korištenje raznih kriterijskih funkcija cilja. Rastave skupa $\mathcal{A}$ na podskupove $\pi_1, \ldots, \pi_k$ koji zadovoljavaju (1) označavat ćemo s $\Pi = \{\pi_1, \ldots, \pi_k\}$ i zvati *particija skupa* $\mathcal{A}$, a skupove $\pi_1, \ldots, \pi_k$ zvat ćemo *klasteri*. Skup svih particija skupa $\mathcal{A}$ sastavljenih od k

*Odjel za matematiku, Sveučilište J. J. Strossmayera u Osijeku, ksabo@mathos.hr
†Odjel za matematiku, Sveučilište J. J. Strossmayera u Osijeku, scitowsk@mathos.hr
‡Odjel za matematiku, Sveučilište J. J. Strossmayera u Osijeku, ivazler@mathos.hr

klastera koje zadovoljavaju (1) označit ćemo s $\mathcal{P}(\mathcal{A}, k)$. Nadalje, kad god budemo govorili o particiji skupa $\mathcal{A}$, podrazumijevat ćemo da je ona sastavljena od podskupova skupa $\mathcal{A}$, koji zadovoljavaju (1). Na taj način svjesno smo iz razmatranja isključili particije, koje sadržavaju prazan skup ili skup $\mathcal{A}$.

Može se pokazati [39] da je broj svih particija skupa $\mathcal{A}$, koje zadovoljavaju (1) jednak Stirlingovom broju druge vrste

$$|\mathcal{P}(\mathcal{A}, k)| = \frac{1}{k!} \sum_{j=1}^{k} (-1)^{k-j} \binom{k}{j} j^m. \tag{2}$$

Primjer 1. *Broj svih particija skupa $\mathcal{A}$ određenih s (1) specijalno za $m = 10, 50, 10^3, 10^6$ i $k = 2, 3, 5, 8, 10$ iznosi*

$\lvert\mathcal{P}(\mathcal{A}, k)\rvert$	$k=2$	$k=3$	$k=5$	$k=8$	$k=10$
$m=10$	511	9330	42525	750	1
$m=50$	10^{15}	10^{23}	10^{33}	10^{40}	10^{43}
$m=10^3$	10^{300}	10^{476}	10^{697}	10^{898}	10^{993}
$m=10^6$	$10^{301\,029}$	$10^{477\,120}$	$10^{698\,968}$	$10^{903\,085}$	10^{10^6}

Iz navedenog primjera vidi se da traženje optimalne particije općenito neće biti moguće provesti pretraživanjem čitavog skupa $\mathcal{P}(\mathcal{A}, k)$. Odmah treba reći da problem traženja optimalne particije spada u NP-teške probleme (vidi [13]) nekonveksne optimizacije općenito nediferencijabilne funkcije više varijabli, koja najčešće posjeduje značajan broj stacionarnih točaka. U znanstvenoj i stručnoj literaturi ovaj problem često nalazimo pod nazivom *cluster analysis*.

Namjera nam je na što jednostavniji način približiti ovo znanstveno područje što širem krugu stručnjaka jer problem klasifikacije i rangiranja podataka u posljednje vrijeme postaje sve zanimljivije područje interesa raznim znanstvenicima i stručnjacima, ali također i donositeljima raznih odluka, primjerice u tijelima državne i lokalne administracije.

Postoji više međunarodno poznatih i priznatih specijaliziranih časopisa koji prate ovo područje, primjerice: *Clustering and Classification*, *Pattern Recognition*, *Journal of Classification*, *Journal of Machine Learning Research*. Također, redovito se održavaju specijalizirani znanstveni i stručni skupovi s ovom problematikom, primjerice: *SIAM International Conference on Data Mining*, *IEEE International Conference on Data Mining*, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, a na brojnim međunarodnim pretraživačima i bazama podataka postoji ogroman broj podataka o ovom području.

U novijoj znanstvenoj i stručnoj literaturi mogu se pronaći brojne primjene u poljoprivredi (primjerice, razvrstavanje oranica prema plodnosti zemljišta), u biologiji (primjerice, klasifikacija kukaca u grupe), u medicini [10, 25], u prometu [17], kod klasifikacije teksta i kompjuterskih pretraživača baza podataka [22, 36], kod razumijevanja klimatskih kretanja i problema lokacije objekata [6, 13], kod satne prognoze potrošnje prirodnog plina i drugih energenata [31]. Također, često se pojavljuju razne primjene u upravljanju (primjerice, rangiranje gradova i općina za potrebe financijske podrške) [15], u poslovanju [10], u društvenim znanostima i psihologiji, kod definiranja izbornih sustava [5, 19] itd.

Posebno naglašavamo da je uz ovaj tekst izrađena i odgovarajuća programska podrška dostupna na `http://www.mathos.hr/oml/software.htm` s ciljem da zainteresirani stručnjaci u svom znanstvenom ili stručnom radu mogu olakšano koristiti ovu metodologiju i pristup.

U sljedećem odjeljku navodimo osnovni pripremni materijal, koji će se koristiti u cijelom radu. U Odjeljku 3. razmatra se najjednostavniji problem grupiranja podataka s jednim obilježjem i to posebno prema principu najmanjih kvadrata i prema principu najmanjih apsolutnih odstupanja. Također u ovom odjeljku navode se i osnovne metode za rješavanje ovog problema za podatke s jednim ili više obilježja. Odjeljak 4. bavi se problemom grupiranja podataka s dva obilježja, a Odjeljak 5. problemom izbora optimalnog broja klastera. U Odjeljku 6. s raznih aspekata analiziramo jedan konkretan problem uz primjenu klaster analize: grupiranje i rangiranje studenata u okviru Bolonjskog procesa studiranja. U posljednjem odjeljku dajemo osnovne karakteristike i upute za korištenje prateće programske podrške.

2. Pripremni materijali

Pretpostavimo da je zadan skup realnih brojeva $A = \{a_1, \ldots, a_m\}$, *među kojima može biti jednakih.* Treba definirati realni broj koji će na neki način reprezentirati taj skup. U tu svrhu najčešće se koristi:

- **aritmetička sredina** $\bar{a} := \frac{1}{m} \sum_{i=1}^{m} a_i$, za koju ističemo sljedeća svojstva

$$\sum_{i=1}^{m} (a_i - \lambda)^2 \geq \sum_{i=1}^{m} (a_i - \bar{a})^2, \qquad \forall \lambda \in \mathbb{R}, \tag{3}$$

$$\sum_{i=1}^{m} (a_i - \bar{a}) = 0; \tag{4}$$

- **medijan** $\mathrm{med}(A)$, za kojeg ističemo sljedeće svojstvo

$$\sum_{i=1}^{m} |a_i - \lambda| \geq \sum_{i=1}^{m} |a_i - \mathrm{med}(A)|, \qquad \forall \lambda \in \mathbb{R}. \tag{5}$$

Aritmetička sredina $\bar{a}$ realnih brojeva $a_1, \ldots, a_m \in \mathbb{R}$ jedinstveni je broj koji ima svojstvo da je suma kvadrata odstupanja brojeva a_i do nekog čvrstog realnog broja najmanja onda ako je taj čvrsti broj upravo aritmetička sredina $\bar{a}$. Princip *najmanje sume kvadrata odstupanja* pripisuje se njemačkom matematičaru C. F. Gaussu[1].

Ako su elementi skupa A sortirani od najmanjeg prema najvećem, onda se medijan može jednostavno zapisati na sljedeći način:

$$\mathrm{med}(A) = \begin{cases} a_{k+1}, & m = 2k+1 \\ \text{bilo koji broj iz segmenta } [a_k, a_{k+1}], & m = 2k \end{cases}. \tag{6}$$

To je takav broj koji ima svojstvo da je suma apsolutnih odstupanja brojeva a_i do nekog čvrstog realnog broja najmanja onda ako je taj čvrsti broj upravo medijan $\mathrm{med}(A)$. Pri tome vrijedi [38]

$$\sum_{i=1}^{m} |a_i - \mathrm{med}(A)| = \sum_{i=1}^{k} (a_{m-i+1} - a_i)\,. \tag{7}$$

[1] Johann Carl Friedrich Gauss (1777-1855)

Princip *najmanje sume apsolutnih odstupanja* pripisuje se hrvatskom znanstveniku J. R. Boškoviću[2]. Više detalja o navedenim pojmovima može se vidjeti kod [30].

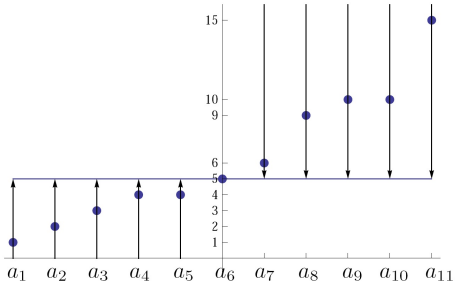
Primjedba 1. *Aritmetička sredina* $\bar{a}$ *skupa podataka - realnih brojeva* $A = \{a_1, \ldots, a_m\}$ *podjednako ovisi o svim podacima. Ako pri tome među podacima ima i onih koji se ekstremno razlikuju od većine podataka, onda će upravo ti ekstremni podaci značajnije utjecati na aritmetičku sredinu. Primijetimo također da se zbog svojstva* (4)*, aritmetička sredina podataka neće promijeniti ako podatke promijenimo tako da je ukupna promjena jednaka nuli.*

Medijan $\mathrm{med}(A)$ *skupa podataka - realnih brojeva* $A = \{a_1, \ldots, a_m\}$ *je srednja veličina na koju ekstremni podaci nemaju nikakav utjecaj. Medijan* $\mathrm{med}(A)$ *neće se promijeniti ako podatke koji su manji od* $\mathrm{med}(A)$ *po volji smanjujemo ili povećavamo do* $\mathrm{med}(A)$*, a podatke koji su veći od* $\mathrm{med}(A)$ *po volji povećavamo ili smanjujemo do* $\mathrm{med}(A)$ *(vidi Sliku 1), tj.*

$$\mathrm{med}(A + E) = \mathrm{med}(A), \tag{8}$$

gdje su $E = \{\delta_1, \ldots, \delta_m\}$ *i* $A + E = \{a_1 + \delta_1, \ldots, a_m + \delta_m\}$ *takvi da je*

$$\begin{array}{lll} (i) & \delta_j \in \{\delta \in \mathbb{R} \colon \delta + a_j > \mathrm{med}(A)\}, & \text{ako je } a_j > \mathrm{med}(A), \\ (ii) & \delta_j \in \{\delta \in \mathbb{R} \colon \delta + a_j < \mathrm{med}(A)\}, & \text{ako je } a_j < \mathrm{med}(A), \\ (iii) & \delta_j = 0, & \text{ako je } a_j = \mathrm{med}(A). \end{array} \tag{9}$$



Slika 1: Područje moguće promjene podataka koja ne utječe na medijan

Najprije primijetimo da je $\mathrm{med}(E) = \delta_\mu = 0$ *i zamislimo da smo skup podataka* $A = \{a_1, \ldots, a_m\}$ *sortirali od najmanjeg prema najvećem. Primjerice, ako je* $\mathrm{med}(A) = a_\mu$*, zbog* (9) *vrijedi*

$$\mathrm{med}(A + E) = \mathrm{med}(\{a_1 + \delta_1, \ldots, a_{\mu-1} + \delta_{\mu-1}, \mathrm{med}(A), a_{\mu+1} + \delta_{\mu+1}, \ldots, a_m + \delta_m\}) = \mathrm{med}(A).$$

Navedena svojstva aritmetičke sredine i medijana imat će vrlo važnu ulogu prilikom grupiranja podataka u skupine u Odjeljcima 3., 4. i 6..

Primjer 2. *Zadan je skup* $A = \{1, 2, 3, 4, 4, 5, 6, 9, 10, 10, 15\}$*. Aritmetička sredina je* $\bar{a} \approx 6.3$*, a medijan* $\mathrm{med}(A) = a_5 = 5$*. Na Slici 1 grafički je prikazano područje moguće promjene podataka koja ne utječe na promjenu medijana.*

Analogno, reprezentant konačnog skupa vektora $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$, *među kojima može biti jednakih*, gdje su $\mathbf{a}_i = (x_i, y_i)$, $i = 1, \ldots, m$, $m \geq 2$ vektori iz $\mathbb{R}^2$, definirat ćemo kao:

[2]Josip Ruđer Bošković (1711-1787)

- **centroid skupa vektora**

$$c(A) = (\bar{x}, \bar{y}) \in \mathbb{R}^2, \qquad \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i, \tag{10}$$

za kojeg ističemo sljedeća svojstva

$$\sum_{i=1}^{m} \|\mathbf{a}_i - \mathbf{u}\|_2^2 \geq \sum_{i=1}^{m} \|\mathbf{a}_i - c(A)\|_2^2, \qquad \forall \mathbf{u} \in \mathbb{R}^2, \tag{11}$$

$$\sum_{i=1}^{m} (\mathbf{a}_i - c(A)) = \mathbf{0}; \tag{12}$$

- **medijan skupa vektora**

$$\mathrm{med}(A) = (\mathrm{med}(\mathbf{x}), \mathrm{med}(\mathbf{y})) \in \mathbb{R}^2, \quad \mathbf{x} = (x_1, \ldots, x_m),\ \mathbf{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m, \tag{13}$$

za kojeg ističemo sljedeće svojstvo

$$\sum_{i=1}^{m} \|\mathbf{a}_i - \mathbf{u}\|_1 \geq \sum_{i=1}^{m} \|\mathbf{a}_i - \mathrm{med}(A)\|_1, \quad \forall \mathbf{u} \in \mathbb{R}^2. \tag{14}$$

Centroid $c(A) = (\bar{x}, \bar{y}) \in \mathbb{R}^2$ skupa vektora $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$, $\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, m$ jedinstveni je vektor koji ima svojstvo da je suma kvadrata euklidskih udaljenosti odgovarajućih točaka (x_i, y_i) do neke čvrste točke iz $\mathbb{R}^2$ najmanja onda ako je ta čvrsta točka upravo $(\bar{x}, \bar{y})$.

Medijan $\mathrm{med}(A) = (\mathrm{med}(x), \mathrm{med}(y)) \in \mathbb{R}^2$ skupa vektora $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$, $\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, m$ je takav vektor koji ima svojstvo da je suma l_1-udaljenosti[3] odgovarajućih točaka $(x_i, y_i) \in \mathbb{R}^2$ do neke čvrste točke iz $\mathbb{R}^2$ najmanja onda ako je ta čvrsta točka upravo $(\mathrm{med}\, x, \mathrm{med}\, y)$. Više detalja o navedenim pojmovima može se vidjeti kod [38].

Primjedba 2. *Pojedine vektore skupa $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$, $\mathbf{a}_i \in \mathbb{R}^2$, $i = 1, \ldots, m$, sukladno Primjedbi 1 moguće je po komponentama mijenjati, a da se centroid, odnosno medijan skupa vektora A ne promijeni.*

Pored centroida i medijana, u literaturi se često razmatra i *geometrijski medijan* $g(A)$ skupa vektora $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$, $\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2$, $i = 1, \ldots, m$, koji se definira kao vektor koji ima svojstvo da je suma euklidskih udaljenosti odgovarajućih točaka $(x_i, y_i) \in \mathbb{R}^2$ do neke čvrste točke iz $\mathbb{R}^2$ najmanja onda ako je ta čvrsta točka upravo geometrijski medijan $g(A)$, tj.

$$\sum_{i=1}^{m} \|\mathbf{a}_i - \mathbf{u}\|_2 \geq \sum_{i=1}^{m} \|\mathbf{a}_i - g(A)\|_2, \qquad \forall \mathbf{u} \in \mathbb{R}^2. \tag{15}$$

Geometrijski medijan skupa vektora ne može se općenito eksplicitno izračunati, a najpoznatiji algoritam za njegovo približno izračunavanje je Weiszfeldov algoritam [16, 19, 33]. U literaturi ovaj problem poznat je pod nazivom *Fermat – Torricelli – Weberov problem* [8, 33].

Primjedba 3. *Analogno bi se mogli definirati centroid, medijan i geometrijski medijan skupa vektora iz $\mathbb{R}^n$.*

[3] U literaturi iz operacijskih istraživanja poznata pod nazivom Manhattan udaljenost [33]

Primijetite da formule (10) – (14) slijede na osnovi sljedeće leme, koja se lako dokazuje [2]. Zbog jednostavnosti, uvedimo oznaku $\underset{x\in\mathcal{D}}{\operatorname{argmin}} h(x)$ za skup svih točaka u kojima funkcija $h : \mathcal{D} \to \mathbb{R}$ postiže globalni minimum. Specijalno, skup $\underset{x\in\mathcal{D}}{\operatorname{argmin}} h(x)$ može biti i jednočlan.

Lema 1. *Neka su $\varphi_i \colon \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$ konveksne funkcije i neka je*

$$\hat{c}_i = \underset{x\in\mathbb{R}}{\operatorname{argmin}} \varphi_i(x), \quad i = 1, \ldots, n.$$

Tada je $f \colon \mathbb{R}^n \to \mathbb{R}$, $f(x_1, \ldots, x_n) = \varphi_1(x_1) + \cdots + \varphi_n(x_n)$ konveksna funkcija i vrijedi

$$\underset{(x_1,\ldots,x_n)\in\mathbb{R}^n}{\operatorname{argmin}} f(x) = (\hat{c}_1, \ldots, \hat{c}_n).$$

3. Grupiranje na osnovi jednog obilježja

Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup koji na osnovi samo jednog obilježja treba grupirati u k klastera koji zadovoljavaju (1). Primjerice, dane u godini možemo grupirati prema prosječnoj dnevnoj temperaturi izraženoj u °C. Svaki element $a_i \in \mathcal{A}$ temeljem tog obilježja reprezentirat ćemo jednim realnim brojem, kojeg ćemo također označavati s a_i. *Zato ćemo nadalje govoriti o skupu podataka-realnih brojeva $\mathcal{A} = \{a_1, \ldots, a_m\}$ među kojima može biti jednakih.*

Funkciju[4] $d \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, koja zadovoljava svojstvo pozitivne definitnosti

$$d(x, y) \geq 0 \quad \forall\, x, y \in \mathbb{R} \qquad \text{i} \qquad d(x, y) = 0 \Leftrightarrow x = y,$$

zvat ćemo *kvazimetrička funkcija.* U literaturi [19, 37] ovakve funkcije nalazimo pod nazivom *"distance like functions".* Kvazimetričke funkcije nalikuju metričkoj funkciji, ali ne moraju zadovoljavati nejednakost trokuta, a u nekim slučajevima nemaju ni svojstvo simetričnosti. U ovom odjeljku koristit ćemo dva tipa kvazimetričkih funkcija:

$$d(a, b) = (a - b)^2, \qquad d(a, b) = |a - b|.$$

Neke druge kvazimetričke funkcije navest ćemo u Odjeljku 4.

Ako je zadana neka kvazimetrička funkcija $d \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar c_j na sljedeći način

$$c_j = c(\pi_j) := \underset{x\in\mathbb{R}}{\operatorname{argmin}} \sum_{a_i\in\pi_j} d(x, a_i). \tag{16}$$

Nadalje, ako na skupu svih particija $\mathcal{P}(\mathcal{A}, k)$ skupa $\mathcal{A}$ sastavljenih od k klastera definiramo kriterijsku funkciju cilja $\mathcal{F} : \mathcal{P}(\mathcal{A}, k) \to \mathbb{R}_+$,

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i\in\pi_j} d(c_j, a_i), \tag{17}$$

[4] U cijelom tekstu skup svih nenegativnih realnih brojeva označavat ćemo s $\mathbb{R}_+$

onda d-optimalnu particiju $\Pi^\star$ tražimo rješavanjem sljedećeg optimizacijskog problema

$$\mathcal{F}(\Pi^\star) = \min_{\Pi \in \mathcal{P}(\mathcal{A},k)} \mathcal{F}(\Pi). \tag{18}$$

Primijetite da na taj način optimalna particija $\Pi^\star$ ima svojstvo da je suma "rasipanja" (suma odstupanja) elemenata klastera oko svog centra minimalna. Na taj način nastojimo postići što bolju unutrašnju kompaktnost i separiranost klastera.

Obrnuto, za dani skup centara $c_1, \ldots, c_k \in \mathbb{R}$, uz primjenu *principa minimalnih udaljenosti* možemo definirati particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ skupa $\mathcal{A}$ na sljedeći način:

$$\pi_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a),\ \forall s = 1, \ldots, k\}, \qquad j = 1, \ldots, k, \tag{19}$$

pri čemu treba voditi računa o tome da svaki element skupa $\mathcal{A}$ pripadne samo jednom klasteru. Zato se problem traženja optimalne particije skupa $\mathcal{A}$ može svesti na sljedeći optimizacijski problem

$$\min_{c_1,\ldots,c_k \in \mathbb{R}} F(c_1, \ldots, c_k), \qquad F(c_1, \ldots, c_k) = \sum_{i=1}^{m} \min_{j=1,\ldots,k} d(c_j, a_i), \tag{20}$$

gdje je $F\colon \mathbb{R}^k \to \mathbb{R}_+$. Općenito, ova funkcija nije konveksna ni diferencijabilna, a može imati više lokalnih minimuma [13, 16, 37]. Optimizacijski problem (20) u literaturi se može naći pod nazivom *k-median problem*, a najčešće se rješava raznim metaheurističkim metodama [7] ili uz primjenu cjelobrojnog programiranja [26, 32, 29]. Pregled radova iz ovog područja do 2006. godine može se vidjeti kod [28].

3.1. Kriterij najmanjih kvadrata

Definicija 1. *Kažemo da je particija $\Pi^\star$ optimalna u smislu najmanjih kvadrata*[5] *(skraćeno: LS-optimalna) ako je $\Pi^\star$ rješenje optimizacijskog problema* (17)–(18)*, a kvazimetrička funkcija $d\colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ definirana s*

$$d(a, b) = (a - b)^2. \tag{21}$$

Primijetite da funkcija (21) nije metrika jer ne zadovoljava nejednakost trokuta. Prema (3), centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$c_j = \operatorname*{argmin}_{u \in \mathbb{R}} \sum_{a_i \in \pi_j} (a_i - u)^2 = \frac{1}{|\pi_j|} \sum_{a_i \in \pi_j} a_i, \qquad j = 1, \ldots, k, \tag{22}$$

a funkcija cilja (17) s

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i \in \pi_j} (c_j - a_i)^2. \tag{23}$$

Primjer 3. *Zadan je skup $\mathcal{A} = \{0, 3, 6, 9\}$. Treba pronaći sve njegove dvočlane particije koje zadovoljavaju* (1)*, odrediti pripadne centre i vrijednosti kriterijske funkcije cilja $\mathcal{F}$ u smislu najmanjih kvadrata.*

Broj svih dvočlanih particija ovog skupa je $2^{m-1} - 1 = 7$, a kao što se vidi iz Tablice 1 LS-optimalna particija u ovom slučaju je $\{\{0, 3\}, \{6, 9\}\}$ jer na njoj kriterijska funkcija cilja $\mathcal{F}$ zadana s (23) postiže najmanju vrijednost.

[5] engl. Least Squares, njem. Varianzkriterium

π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$		$\mathcal{G}(\Pi)$	
{0}	{3,6,9}	0	6	0+18	=18	81/4+27/4	= 27
{3}	{0,6,9}	3	5	0+42	=42	9/4+3/4	= 3
{6}	{0,3,9}	6	4	0+42	=42	9/4+3/4	= 3
{9}	{0,3,6}	9	3	0+18	=18	81/4+27/4	= 27
{0,3}	{6,9}	3/2	15/2	9/2+9/2	=9	18+18	= 36
{0,6}	{3,9}	3	6	18+18	=36	9/2+9/2	= 9
{0,9}	{3,6}	9/2	9/2	81/2+9/2	=45	0+0	= 0

Tablica 1: Particije, centri i funkcije cilja $\mathcal{F}$ i $\mathcal{G}$

3.1.1. Dualni problem

Sljedeća lema pokazuje da je "rasipanje" skupa $\mathcal{A}$ oko njegovog centra c jednako zbroju "rasipanja" klastera π_j, $j = 1, \ldots, k$, oko njihovih centara c_j, $j = 1, \ldots, k$, i težinskoj sumi kvadrata odstupanja centra c od centara c_j, pri čemu su težine određene veličinom skupova π_j.

Lema 2. *Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup podataka, a $\Pi = \{\pi_1, \ldots, \pi_k\}$ neka particija s klasterima $\pi_1, \ldots, \pi_k$ duljine $m_1, \ldots, m_k$. Neka je nadalje*

$$c = \frac{1}{m}\sum_{i=1}^{m} a_i, \qquad c_j = \frac{1}{m_j}\sum_{a_i \in \pi_j} a_i, \quad j = 1, \ldots, k. \tag{24}$$

Tada vrijedi

$$\sum_{i=1}^{m}(a_i - c)^2 = \mathcal{F}(\Pi) + \mathcal{G}(\Pi), \tag{25}$$

gdje je

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k}\sum_{a_i \in \pi_j}(c_j - a_i)^2, \tag{26}$$

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j (c_j - c)^2. \tag{27}$$

Dokaz. Primijetimo najprije da za svaki $x \in \mathbb{R}$ vrijedi

$$\sum_{a_i \in \pi_j}(a_i - x)^2 = \sum_{a_i \in \pi_j}(a_i - c_j)^2 + m_j(c_j - x)^2, \quad j = 1, \ldots, k. \tag{28}$$

Naime, kako je prema (4), $\sum\limits_{a_i \in \pi_j}(a_i - c_j)(c_j - x) = (c_j - x)\sum\limits_{a_i \in \pi_j}(a_i - c_j) = 0$, vrijedi

$$\begin{aligned}\sum_{a_i \in \pi_j}(a_i - x)^2 &= \sum_{a_i \in \pi_j}((a_i - c_j) + (c_j - x))^2 \\ &= \sum_{a_i \in \pi_j}(a_i - c_j)^2 + m_j(c_j - x)^2.\end{aligned}$$

Ako u (28) umjesto x stavimo $c = \frac{1}{m}\sum_{i=1}^{m} a_i$ i zbrojimo sve jednakosti, dobivamo (25). □

Iz Leme 2 neposredno slijedi tvrdnja sljedećeg teorema [5, 35]

Teorem 1. *Uz oznake kao u Lemi 2 vrijedi:*

$$\operatorname*{argmin}_{\Pi\in\mathcal{P}(\mathcal{A},k)} \mathcal{F}(\Pi) = \operatorname*{argmax}_{\Pi\in\mathcal{P}(\mathcal{A},k)} \mathcal{G}(\Pi).$$

To znači da u cilju pronalaženja LS-optimalne particije, umjesto minimizacije funkcije $\mathcal{F}$ zadane s (23), odnosno (26), možemo maksimizirati funkciju

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j (c_j - c)^2. \tag{29}$$

Primjer 4. *Centar skupa $\mathcal{A} = \{0, 3, 6, 9\}$ iz Primjera 3 je $c(\mathcal{A}) = \frac{9}{2}$, a ukupno rasipanje skupa $\mathcal{A}$ oko centra $c(\mathcal{A})$ je $\sum_{i=1}^{m}(a_i - c)^2 = 45$. Za svaku od 7 različitih particija u Tablici 1 prikazana je također i vrijednost kriterijske funkcije cilja $\mathcal{G}$. Kao što se vidi, funkcija $\mathcal{G}$ prima najveću vrijednost na optimalnoj particiji $\{\{0, 3\}, \{6, 9\}\}$, što je u skladu s Teoremom 1. Također, u skladu s Lemom 2, za svaku particiju Π vrijedi $\mathcal{F}(\Pi) + \mathcal{G}(\Pi) = 45$ (vidi također Tablicu 1).*

Primjedba 4. *Lako se može provjeriti da je veza između centra c čitavog skupa $\mathcal{A}$ i centara c_j pojedinih klastera π_j zadanih s (24) dana s*

$$c = \frac{m_1}{m} c_1 + \cdots + \frac{m_k}{m} c_k.$$

Specijalno, za dva disjunktna skupa realnih brojeva $A = \{x_1, \ldots, x_p\}$, $B = \{y_1, \ldots, y_q\}$ aritmetička sredina njihove unije jednaka je ponderiranom zbroju njihovih aritmetičkih sredina, tj. vrijedi

$$\overline{A \cup B} = \frac{p}{p+q}\overline{A} + \frac{q}{p+q}\overline{B}.$$

3.2. Kriterij najmanjih apsolutnih odstupanja

Definicija 2. *Kažemo da je particija $\Pi^\star$ optimalna u smislu najmanjih apsolutnih odstupanja[6] (skraćeno: LAD-optimalna) ako je $\Pi^\star$ rješenje optimizacijskog problema (17)–(18), a metrička funkcija $d\colon \mathbb{R}\times\mathbb{R} \to \mathbb{R}_+$ definirana s*

$$d(a, b) = |a - b|. \tag{30}$$

Prema (5) centri $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$c_j = \operatorname*{argmin}_{u\in\mathbb{R}} \sum_{a_i\in\pi_j} |a_i - u| = \operatorname{med}(\pi_j), \qquad j = 1, \ldots, k, \tag{31}$$

a funkcija cilja (17) s

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i\in\pi_j} |c_j - a_i|. \tag{32}$$

Ako pri tome iskoristimo (7), onda za izračunavanje funkcije cilja (32) nije potrebno poznavati centre klastera (31), što može značajno ubrzati računski proces.

[6] engl. Least Absolute Deviations

3.3. Grupiranje podataka s težinama

Pretpostavimo da je zadan skup podataka $\mathcal{A} = \{a_1, \ldots, a_m\}$, pri čemu je svakom podatku a_i pridužena odgovarajuća težina $w_i > 0$. Kriterijska funkcija cilja (17) sada postaje

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{a_i \in \pi_j} w_i d(c_j, a_i). \tag{33}$$

Specijalno, kod primjene kriterija LS-optimalnosti centar c_j klastera π_j određen je težinskom aritmetičkom sredinom podataka iz klastera π_j

$$c_j = \frac{1}{\kappa_j} \sum_{a_i \in \pi_j} w_i a_i, \qquad \kappa_j = \sum_{a_i \in \pi_j} w_i, \tag{34}$$

a kod primjene kriterija LAD-optimalnosti centar c_j klastera π_j određen je težinskim medijanom podataka iz klastera π_j (vidi [30, 38]):

$$c_j = \operatorname*{med}_{a_i \in \pi_j} (w_i, a_i). \tag{35}$$

3.4. Problem traženja optimalne particije je problem globalne optimizacije

Kao što je već ranije spomenuto, problem traženja optimalne particije skupa $\mathcal{A}$ je problem globalne optimizacije, a minimizirajuća funkcija F definirana s (20) nije ni konveksna ni diferencijabilna, a može imati više lokalnih minimuma. U praktičnim primjerima pokazuje se da već kod dvadesetak podataka i pet klastera (vidi Primjer 10) taj broj lokalnih minimuma može biti neočekivano velik.

Problem traženja globalnog minimuma funkcija više varijabli općenito je vrlo složen problem. Pregled radova iz ovog područja objavljenih u posljednje vrijeme može se naći kod [12]. U knjizi [14] dan je pristup rješavanju ovog problema preko tzv. *intervalne analize*. Na osnovi radova [27, 34] izrađen je efikasan algoritam globalne optimizacije za tzv. klasu Lipschitzovih funkcija nazvan `DIRECT` [11, 18]. Algoritam `DIRECT` može se vrlo uspješno primijeniti za rješavanje problema traženja optimalne particije.

Budući da u našem slučaju minimizirajuća funkcija (20) nije diferencijabilna, problem postaje još složeniji. Ako bi rješenje pokušali direktno dobiti pretraživanjem svih mogućih particija, to bi računski bilo vrlo zahtjevno i iziskivalo bi značajno vrijeme rada računala: u slučaju većeg broja podataka i klastera to postaje gotovo nemoguć pothvat (vidi Primjer 1).

3.4.1. Standardni k-means algoritam

Uz pretpostavku da smo na neki način dobro procijenili početnu aproksimaciju centara klastera ili dobru početnu particiju, niže navedenim algoritmom možemo dobiti particiju dosta blisku optimalnoj [19, 37]. Algoritam ćemo napisati dovoljno općenito uz korištenje kvazimetričke funkcije $d\colon X \times X \to \mathbb{R}_+$, gdje je X prostor podataka (vidi [24]).

Algoritam 1. ***(Standardni k-means algoritam)***

Korak 0: *Učitati m, k, skup $\mathcal{A}$ i izabrati početne centre $c_1^0, \ldots, c_k^0$;*

Korak 1: *Primjenom principa minimalnih udaljenosti odrediti početnu particiju $\Pi = \{\pi_1, \ldots, \pi_k\}$ tako da neki $a \in \mathcal{A}$ pripadne onom klasteru čiji je centar najbliži elementu a. Izračunati centre $c_1, \ldots, c_k$ klastera $\pi_1, \ldots, \pi_k$ i početnu vrijednost funkcije cilja $F_0 = \mathcal{F}(\Pi)$;*

Korak 2: *Formirati novu particiju $\mathcal{N} = \{\nu_1, \ldots, \nu_k\}$ tako da neki $a \in \mathcal{A}$ pripadne onom klasteru čiji je centar najbliži elementu a, njihove centroide $\zeta_1, \ldots, \zeta_k$ i novu vrijednost funkcije cilja $F_1 = \mathcal{F}(\mathcal{N})$;*

Korak 3: *Ako je $F_1 < F_0$, staviti $c_j = \zeta_j$; $j = 1, \ldots, k$; $F_0 = F_1$ i prijeći na Korak 2;*
U protivnom, STOP.

Primjedba 5. *Za podatke s jednim obilježjem u slučaju izbora LAD-kriterija optimalnosti metrička funkcija d zadana je s* (30). *U tom slučaju u Koraku 2 Algoritma 1 može se dogoditi da neki centroid ζ_j, sukladno* (6), *može biti proizvoljan broj iz nekog intervala $[\alpha, \beta] \subset \mathbb{R}$. U tom slučaju treba uzeti $\zeta_j = \frac{\alpha+\beta}{2}$.*

3.4.2. Traženje optimalne particije na osnovi jednog obilježja

Problem traženja optimalne k-člane particije skupa $\mathcal{A} \subset \mathbb{R}$ s jednim obilježjem nešto je jednostavniji, iako se i u ovom slučaju općenito radi o optimizacijskom problemu za nekonveksnu i/ili nediferencijabilnu funkciju više varijabli.

(i) Specijalno, u ovom slučaju princip minimalnih udaljenosti kojim se na osnovi zadanih centara c_j, $j = 1, \ldots, k$, određuju odgovarajuće particije π_j, (Korak 1 i Korak 2) u Algoritmu 1, ne ovisi o izboru kvazimetričke funkcije d. Uočimo da tada u Koraku 1 Algoritma 1 možemo pisati

$$\pi_1 = \{a \in \mathcal{A} : a \leq \frac{1}{2}(c_1^0 + c_2^0)\},$$
$$\pi_j = \{a \in \mathcal{A} : \frac{1}{2}(c_{j-1}^0 + c_j^0) < a \leq \frac{1}{2}(c_j^0 + c_{j+1}^0)\}, \quad j = 2, \ldots, k-1,$$
$$\pi_k = \{a \in \mathcal{A} : a > \frac{1}{2}(c_{k-1}^0 + c_k^0)\},$$

dok u Koraku 2, možemo pisati

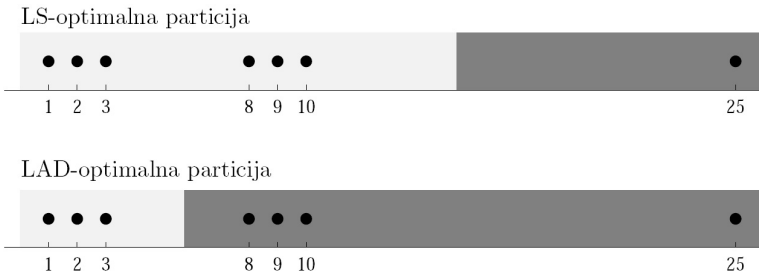
$$\nu_1 = \{a \in \mathcal{A} : a \leq \frac{1}{2}(c_1 + c_2)\},$$
$$\nu_j = \{a \in \mathcal{A} : \frac{1}{2}(c_{j-1} + c_j) < a \leq \frac{1}{2}(c_j + c_{j+1})\}, \quad j = 2, \ldots, k-1,$$
$$\nu_k = \{a \in \mathcal{A} : a > \frac{1}{2}(c_{k-1} + c_k)\}.$$

Primjer 5. *Zadan je skup $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$. Primjenom Algoritma 1 treba pronaći dvočlanu particiju što bližu LS-optimalnoj.*

Iteracija	Klasteri	Centri		Funkcija cilja
1.	{1,2,3,8}, {9,10,25}	3.50	14.67	189.67
2.	{1,2,3,8,9}, {10,25}	4.60	17.50	165.70
3.	{1,2,3,8,9,10}, {25}	5.50	25.00	77.50
4.	{1,2,3,8,9,10}, {25}	5.50	25.00	77.50

Tablica 2: Tijek iterativnog postupka Algoritma 1

Broj svih dvočlanih particija ovog skupa je $2^{m-1}-1 = 63$. Odmah uočavamo da skup $\mathcal{A}$ sadrži dvije značajno različite skupine realnih brojeva $\mathcal{A}_1 = \{1, 2, 3\}$ te $\mathcal{A}_2 = \{8, 9, 10\}$. Također, skup $\mathcal{A}$ sadrži i element 25, kojeg možemo shvatiti kao jako stršeći podatak nastao zbog određene pogreške, a prirodno dolazi iz skupine $\mathcal{A}_2$. Primjenom Algoritma 1 uz početne centre $c_1 = 2$ i $c_2 = 15$, dobivamo početnu particiju $\Pi = \{\pi_1, \pi_2\}$, $\pi_1 = \{1, 2, 3, 8\}$, $\pi_2 = \{9, 10, 25\}$. U Tablici 2 prikazan je tijek iterativnog postupka. Direktnom provjerom svih particija može se pokazati da je Algoritam 1 pronašao upravo optimalnu particiju. Iz ovog primjera vidljivo je da k-means algoritam u smislu LS-optimalnosti sukladno Primjedbi 1 daje particiju, koja značajno ovisi o stršećem podatku, tako da upravo stršeći podatak čini zaseban klaster (vidi Sliku 2).



Slika 2: Optimalne particije skupa $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$ dobivene Algoritmom 1

Primjer 6. *Zadan je skup $\mathcal{A} = \{1, 2, 3, 8, 9, 10, 25\}$ iz Primjera 5. Primjenom Algoritma 1 treba pronaći dvočlanu particiju skupa $\mathcal{A}$ što bližu LAD-optimalnoj.*

Iteracija	Klasteri	Centri		Funkcija cilja
1.	{1,2,3,8}, {9,10,25}	2.50	10.00	24
2.	{1,2,3}, {8,9,10,25}	2.00	9.50	20
3.	{1,2,3}, {8,9,10,25}	2.00	9.50	20

Tablica 3: Tijek iterativnog postupka Algoritma 1

Primjenom Algoritma 1 uz početne centre kao u Primjeru 5, $c_1 = 2$ i $c_2 = 15$, dobivamo početnu particiju $\Pi = \{\pi_1, \pi_2\}$, $\pi_1 = \{1, 2, 3, 8\}$, $\pi_2 = \{9, 10, 25\}$. U Tablici 3 prikazan je tijek iterativnog postupka. Pri tome, centri u Koraku 2 Algoritma 1 birani su u skladu s Primjedbom 5. Direktnom provjerom može se pokazati da je algoritam pronašao upravo

LAD-optimalnu particiju. U ovom slučaju stršeći podatak sukladno Primjedbi 1 prirodno je pridružen drugom klasteru (vidi Sliku 2).

(ii) Nadalje, očigledno je da je optimalnu k-članu particiju sortiranog skupa $\mathcal{A} \subset \mathbb{R}$ s jednim obilježjem dovoljno tražiti između particija $\Pi = \{\pi_1, \ldots, \pi_k\}$, čiji se klasteri nastavljaju jedan na drugi, tj. između particija za čije klastere vrijedi: $\max \pi_i < \min \pi_{i+1}$, $i = 1, \ldots, k-1$. Broj svih takvih particija je $\binom{m-1}{k-1}$ i znatno je manji u usporedbi s brojem svih mogućih particija (2) koje zadovoljavaju samo (1). Niže navednu tablicu usporedite s tablicom iz Primjera 1.

$\binom{m-1}{k-1}$	$k=2$	$k=3$	$k=4$	$k=5$	$k=10$
$m=10$	9	36	84	126	1
$m=20$	19	171	969	3876	92378
$m=50$	49	1176	18424	211876	$\approx 2 \times 10^8$
$m=100$	99	4851	156849	$\approx 4 \times 10^6$	$\approx 2 \times 10^{12}$

U slučaju kada je broj $\binom{m-1}{k-1}$ relativno malen, optimalnu particiju možemo potražiti izračunavanjem vrijednosti funkcije cilja na svim ovakvim particijama (vidi Tablicu 7 u Odjeljku 7). Ako u tom slučaju koristimo kriterij LAD-optimalnosti, za izračunavanje vrijednosti funkcije cilja nije potrebno poznavati centre klastera (vidi Odjeljak 3.2.), što dodatno ubrzava računski proces.

Ako je broj $\binom{m-1}{k-1}$ relativno velik, morat ćemo se zadovoljiti nekom stacionarnom točkom u kojoj funkcija cilja možda neće postići globalni minimum. U tom slučaju između ovih $\binom{m-1}{k-1}$ particija na neki način treba izabrati razuman broj particija, koje će nam poslužiti kao početne particije u Algoritmu 1. Tako dobivena particija s najnižom vrijednosti funkcije cilja može zamijeniti traženu optimalnu particiju.

(iii) Za relativno veliki skup $\mathcal{A} \subset \mathbb{R}$ s jednim obilježjem navest ćemo jedan način izbora početnih centara u Algoritmu 1 koji često dovodi do optimalne particije. Pretpostavimo dakle, da je zadan sortirani skup podataka $A = \{a_1, \ldots, a_m\}$, za koji treba pronaći k-članu LS ili LAD-optimalnu particiju.

Najprije ćemo skup $\mathcal{A}$ razdijeliti na k približno jednakih podskupova $\pi_1, \ldots, \pi_k$ zadržavajući pri tome sortirani redoslijed elemenata. Početne centre $c_1^0, \ldots, c_k^0$ u *Koraku 0* Algoritma 1 odredit ćemo na sljedeći način:

(a) za traženje LS-optimalne particije za c_j^0 treba uzeti aritmetičku sredinu skupa π_j;

(b) za traženje LAD-optimalne particije za c_j^0 treba uzeti medijan skupa π_j.

Ako je zadan skup $\mathcal{A}$ s odgovarajućim težinama $w_i > 0$, onda najprije odredimo najmanji prirodni broj $p \in \mathbb{N}$, takav da je $10^p w_i \geq 1$, $\forall i = 1, \ldots, m$ i definiramo nove težine kao najveći cijeli broj manji ili jednak $10^p w_i$, tj. $\kappa_i = \lfloor 10^p w_i \rfloor \in \mathbb{N}$. Nakon toga definiramo nove podatke

$$\underbrace{a_1, \ldots, a_1}_{\kappa_1}, \underbrace{a_2, \ldots, a_2}_{\kappa_2}, \ldots, \underbrace{a_m, \ldots, a_m}_{\kappa_m},$$

i na njih primijenimo postupak (a), odnosno (b).

4. Grupiranje na osnovi dva obilježja

Neka je $\mathcal{A} = \{a_1, \ldots, a_m\}$ skup koji treba na osnovi dva obilježja grupirati u k klastera koji zadovoljavaju (1). Primjerice, dane u godini možemo grupirati prema prosječnoj dnevnoj temperaturi izraženoj u °C i količini dnevnih padavina. Svaki element $a_i \in \mathcal{A}$ temeljem tih obilježja reprezentirat ćemo jednim vektorom $(x_i, y_i) \in \mathbb{R}^2$, kojeg ćemo označiti s $\mathbf{a}_i$. Nadalje, zbog jednostavnosti elemente skupa $\mathcal{A}$ identificirat ćemo s tim vektorima i govoriti o *skupu podataka-vektora među kojima može biti jednakih.*

Ako je zadana neka kvazimetrička funkcija $d\colon \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$, onda svakom klasteru $\pi_j \in \Pi$ možemo pridružiti njegov centar $\mathbf{c}_j$ na sljedeći način

$$\mathbf{c}_j = c(\pi_j) := \operatorname*{argmin}_{\mathbf{x}\in\mathbb{R}^2} \sum_{\mathbf{a}_i\in\pi_j} d(\mathbf{x}, \mathbf{a}_i). \tag{36}$$

Kod problema grupiranja podataka u općem slučaju najčešće korištene kvazimetričke funkcije $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ su [1, 13, 19]

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2 \qquad (\textit{least squares} \text{ udaljenost}) \tag{37}$$

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 \qquad (\text{Manhattan udaljenost}) \tag{38}$$

$$d(\mathbf{x}, \mathbf{y}) = \frac{1}{2}(\mathbf{x} - \mathbf{y})\mathbf{Q}(\mathbf{x} - \mathbf{y})^T, \quad \mathbf{Q} > 0,\ \mathbf{Q}^T = \mathbf{Q} \qquad (\text{Mahalanobisova udaljenost})$$

$$d(\mathbf{x}, \mathbf{y}) = \sum_i \mathbf{x}_i \left(\ln \frac{\mathbf{x}_i}{\mathbf{y}_i} - \mathbf{x}_i + \mathbf{y}_i \right),\ \mathbf{x}, \mathbf{y} \in \mathbb{R}_+^n\ (0 \ln 0 := 0) \quad (\text{Kullback–Leiblerova udaljenost})$$

Na skupu svih particija $\mathcal{P}(\mathcal{A}, k)$ skupa $\mathcal{A}$ sastavljenih od k klastera, potpuno analogno kao i ranije, definiramo kriterijsku funkciju cilja $\mathcal{F} : \mathcal{P}(\mathcal{A}, k) \to \mathbb{R}_+$,

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{\mathbf{a}_i\in\pi_j} d(\mathbf{c}_j, \mathbf{a}_i), \tag{39}$$

a d-optimalnu particiju $\Pi^\star$ tražimo rješavanjem optimizacijskog problema

$$\mathcal{F}(\Pi^\star) = \min_{\Pi\in\mathcal{P}(\mathcal{A},k)} \mathcal{F}(\Pi). \tag{40}$$

I u ovom slučaju problem traženja optimalne particije može se preformulirati na problem traženja optimalnih centara, pri tome funkcija cilja (39) također može imati više lokalnih minimuma, koje također možemo tražiti primjenom Algoritma 1.

Primijetite da na taj način optimalna particija $\Pi^\star$ ima svojstvo da je suma "rasipanja" (suma odstupanja) elemenata klastera oko svog centra minimalna. Na taj način nastojimo postići što bolju unutrašnju kompaktnost i separiranost klastera.

4.1. Kriterij najmanjih kvadrata

Definicija 3. *Neka je* $\mathcal{A} = \{\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, m\}$ *skup vektora iz* $\mathbb{R}^2$. *Kažemo da je particija* $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ *optimalna u smislu najmanjih kvadrata (skraćeno: LS-optimalna) ako je* $\Pi^\star$ *rješenje optimizacijskog problema* (39)–(40)*, a kvazimetrička funkcija* $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ *definirana s* (37).

Primijetite da funkcija (37) nije metrika jer ne zadovoljava nejednakost trokuta. Prema (10) centri $\mathbf{c}_1, \ldots, \mathbf{c}_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$\mathbf{c}_j = \operatorname*{argmin}_{\mathbf{u}\in\mathbb{R}^2} \sum_{\mathbf{a}_i\in\pi_j} \|\mathbf{a}_i - \mathbf{u}\|_2^2 = \frac{1}{|\pi_j|} \sum_{\mathbf{a}_i\in\pi_j} \mathbf{a}_i, \quad j = 1, \ldots, k, \tag{41}$$

a funkcija cilja (39) s

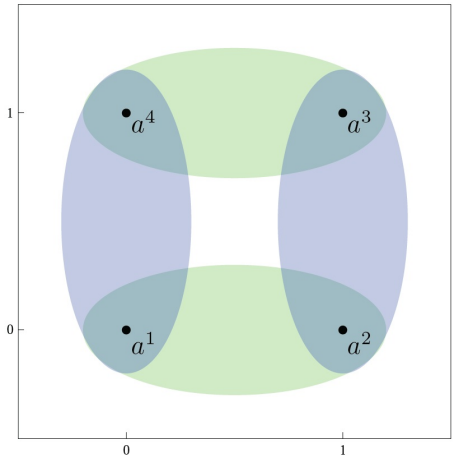
$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{\mathbf{a}_i\in\pi_j} \|\mathbf{c}_j - \mathbf{a}_i\|_2^2 \tag{42}$$

Primjer 7. *Za skup* $\mathcal{A} = \{\mathbf{a}_1 = (0,0), \mathbf{a}_2 = (1,0), \mathbf{a}_3 = (1,1), \mathbf{a}_4 = (0,1)\}$ *odredit ćemo sve dvočlane particije, koje zadovoljavaju* (1)*, a nakon toga odgovarajuće centroide i vrijednosti funkcije cilja* (42) *u smislu LS-optimalnosti.*

Broj svih dvočlanih particija ovog skupa je $2^{m-1} - 1 = 7$, a kao što se vidi iz Tablice 4, dvije particije $\{\{\mathbf{a}_1, \mathbf{a}_2\}, \{\mathbf{a}_3, \mathbf{a}_4\}\}$ i $\{\{\mathbf{a}_1, \mathbf{a}_4\}, \{\mathbf{a}_2, \mathbf{a}_3\}\}$ su optimalne jer na njima kriterijska funkcija cilja (42) postiže globalni minimum (vidi Sliku 3).

π_1	π_2	$\mathbf{c}_1$	$\mathbf{c}_2$	$\mathcal{F}(\Pi)$		$\mathcal{G}(\Pi)$	
$\{\mathbf{a}_1\}$	$\{\mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4\}$	$\mathbf{a}_1$	$\left(\frac{2}{3}, \frac{2}{3}\right)$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6}$	≈ 0.6
$\{\mathbf{a}_2\}$	$\{\mathbf{a}_1, \mathbf{a}_3, \mathbf{a}_4\}$	$\mathbf{a}_2$	$\left(\frac{1}{3}, \frac{2}{3}\right)$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6}$	≈ 0.6
$\{\mathbf{a}_3\}$	$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_4\}$	$\mathbf{a}_3$	$\left(\frac{1}{3}, \frac{1}{3}\right)$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6}$	≈ 0.6
$\{\mathbf{a}_4\}$	$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$	$\mathbf{a}_4$	$\left(\frac{2}{3}, \frac{1}{3}\right)$	$0 + \frac{4}{3}$	≈ 1.3	$\frac{1}{2} + \frac{1}{6}$	≈ 0.6
$\{\mathbf{a}_1, \mathbf{a}_2\}$	$\{\mathbf{a}_3, \mathbf{a}_4\}$	$\left(\frac{1}{2}, 0\right)$	$\left(\frac{1}{2}, 1\right)$	$\frac{1}{2} + \frac{1}{2}$	$= 1$	$\frac{1}{2} + \frac{1}{2}$	$= 1$
$\{\mathbf{a}_1, \mathbf{a}_4\}$	$\{\mathbf{a}_2, \mathbf{a}_3\}$	$\left(0, \frac{1}{2}\right)$	$\left(1, \frac{1}{2}\right)$	$\frac{1}{2} + \frac{1}{2}$	$= 1$	$\frac{1}{2} + \frac{1}{2}$	$= 1$
$\{\mathbf{a}_1, \mathbf{a}_3\}$	$\{\mathbf{a}_2, \mathbf{a}_4\}$	$\left(\frac{1}{2}, \frac{1}{2}\right)$	$\left(\frac{1}{2}, \frac{1}{2}\right)$	$1 + 1$	$= 2$	$0 + 0$	$= 0$

Tablica 4: Particije, centri i funkcije cilja $\mathcal{F}$ i $\mathcal{G}$



Slika 3: LS-optimalne particije skupa $\mathcal{A}$ iz Primjera 7

4.1.1. Dualni problem

Analogno, kao u jednodimenzionalnom slučaju može se pokazati da vrijedi

$$\sum_{i=1}^{m} \|\mathbf{a}_i - \mathbf{c}\|_2^2 = \sum_{j=1}^{k} \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{c}_j - \mathbf{a}_i\|_2^2 + \sum_{j=1}^{k} m_j \|\mathbf{c}_j - \mathbf{c}\|_2^2, \tag{43}$$

gdje je $\mathbf{c} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{a}_i$ centar skupa $\mathcal{A}$, a $m_j = |\pi_j|$. Zato umjesto minimizacije funkcije $\mathcal{F}$ zadane s (42) optimalnu LS-particiju možemo tražiti maksimizacijom funkcije

$$\mathcal{G}(\Pi) = \sum_{j=1}^{k} m_j \|\mathbf{c}_j - \mathbf{c}\|_2^2. \tag{44}$$

Određenim prilagođavanjem [5] problem se svodi na poznate probleme i metode linearne algebre.

Primjer 8. *Skup $\mathcal{A}$ iz Primjera 7 ima 7 različitih particija i za sve njih u Tablici 4 prikazana je vrijednost kriterijske funkcije cilja $\mathcal{G}$. Kao što se vidi iz Tablice 4, funkcija $\mathcal{G}$ prima maksimalnu vrijednost na već ranije dobivenim optimalnim particijama $\{\{\mathbf{a}_1, \mathbf{a}_2\}, \{\mathbf{a}_3, \mathbf{a}_4\}\}$ i $\{\{\mathbf{a}_1, \mathbf{a}_4\}, \{\mathbf{a}_2, \mathbf{a}_3\}\}$.*

4.2. Kriterij najmanjih apsolutnih odstupanja

Definicija 4. *Neka je $\mathcal{A} = \{\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \ldots, m\}$ skup vektora iz $\mathbb{R}^2$. Kažemo da je particija $\Pi^\star = \{\pi_1^\star, \ldots, \pi_k^\star\}$ optimalna u smislu najmanjih apsolutnih odstupanja (skraćeno: LAD-optimalna) ako je $\Pi^\star$ rješenje optimizacijskog problema (39)–(40), a metrička funkcija $d : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}_+$ definirana s (38).*

Prema (13) centri $\mathbf{c}_1, \ldots, \mathbf{c}_k$ klastera $\pi_1, \ldots, \pi_k$ određeni su s

$$\mathbf{c}_j = \operatorname*{argmin}_{\mathbf{u} \in \mathbb{R}^2} \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{a}_i - \mathbf{u}\|_1 = \operatorname{med}(\pi_j), \quad j = 1, \ldots, k, \tag{45}$$

te funkcija cilja (39) zadana je s

$$\mathcal{F}(\Pi) = \sum_{j=1}^{k} \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{c}_j - \mathbf{a}_i\|_1 \tag{46}$$

Primjedba 6. *Kao što smo u Odjeljku 3.3. razmatrali problem grupiranja jednodimenzionalnih težinskih podataka, slično bi mogli postupiti i u slučaju grupiranja težinskih dvodimenzionalnih i višedimenzionalnih podataka [38].*

5. Problem izbora optimalnog broja klastera

U nekim slučajevima (vidi Primjere 9,10) broj klastera u particiji skupa $\mathcal{A}$ određen je iz prirode problema. Ako broj klastera u particiji nije unaprijed zadan, onda je prirodno tražiti optimalnu particiju s klasterima koji su što kompaktnije grupirani i što bolje separirani. U tom smislu može se postaviti i problem optimalnog broja klastera u particiji. O ovom problemu može se naći vrlo različitih pristupa u literaturi (vidi primjerice [10, 13, 17]), a ovdje navodimo nekoliko najčešće korištenih.

(i) Jedna mogućnost je promatrati optimalnu vrijednost kriterijske funkcije cilja kao funkciju broja klastera. Budući da s porastom broja klastera optimalna vrijednost kriterijske funkcije cilja opada [35], kao optimalnu vrijednost broja klastera k možemo uzeti onu vrijednost za koju je kriterijska funkcija cilja naglo pala.

(ii) (Davies – Bouldinov indeks [4]) Za optimalnu particiju s k klastera definiramo indeks

$$V_{DB} = \frac{1}{k}\sum_{i=1}^{k} R_i, \tag{47}$$

gdje je

$$\begin{aligned} R_i &= \max_{j\neq i} R_{ij}, \\ R_{ij} &= \frac{S_i + S_j}{D_{ij}} \quad \text{(mjera sličnosti klastera } \pi_i \text{ i } \pi_j\text{)}, \\ D_{ij} &= d(c_i, c_j) \quad \text{(udaljenost centara klastera } \pi_i \text{ i } \pi_j\text{)}, \\ S_i &= \frac{1}{|\pi_i|}\sum_{a\in\pi_i} d(a, c_i) \quad \text{(rasipanje klastera } \pi_i\text{)}. \end{aligned}$$

Kompaktniji i bolje separirani klasteri u optimalnoj particiji rezultirat će manjim V_{DB} indeksom.

(iii) (Dunnov indeks [9]) Za optimalnu particiju s k klastera definiramo indeks

$$V_D = \min_{1\leq i<j\leq k}\left(\frac{D(\pi_i,\pi_j)}{\max\limits_{1\leq s\leq k}\operatorname{diam}\pi_s}\right), \tag{48}$$

gdje je

$$\begin{aligned} D(\pi_i,\pi_j) &= \min_{a\in\pi_i,\, b\in\pi_j} d(a,b), \\ \operatorname{diam}\pi_i &= \max_{a,b\in\pi_i} d(a,b). \end{aligned}$$

Kompaktniji i bolje separirani klasteri u optimalnoj particiji rezultirat će manjim V_D indeksom.

(iv) (Calinski-Harabaszev indeks [3, 10]) U slučaju primjene kriterija LS-optimalnosti možemo koristiti Calinski-Harabaszev indeks

$$V_{CH} = \frac{(m-k)\mathcal{G}(\mathcal{A})}{(k-1)\mathcal{F}(\mathcal{A})}, \tag{49}$$

gdje su $\mathcal{F}(\mathcal{A})$, odnosno $\mathcal{G}(\mathcal{A})$, funkcije zadane s (26), odnosno s (27). Kompaktniji i bolje separirani klasteri u optimalnoj particiji rezultirat će većim V_{CH} indeksom.

6. Mjerenje uspješnosti studenata i Bolonjski proces

Iako postoje brojne mogućnosti primjene klaster analize navedene u Uvodu, odlučili smo se ovu metodu analize podataka ilustrirati na aktualnim problemima i primjenama iz obrazovnog sustava. S tim u vezi razmatramo problem grupiranja studenata prema rezultatima postignutim u okviru jednog predmeta i problem rangiranja studenata na osnovi prosječne ocjene i postignutog broja ECTS bodova u jednoj studijskoj godini.

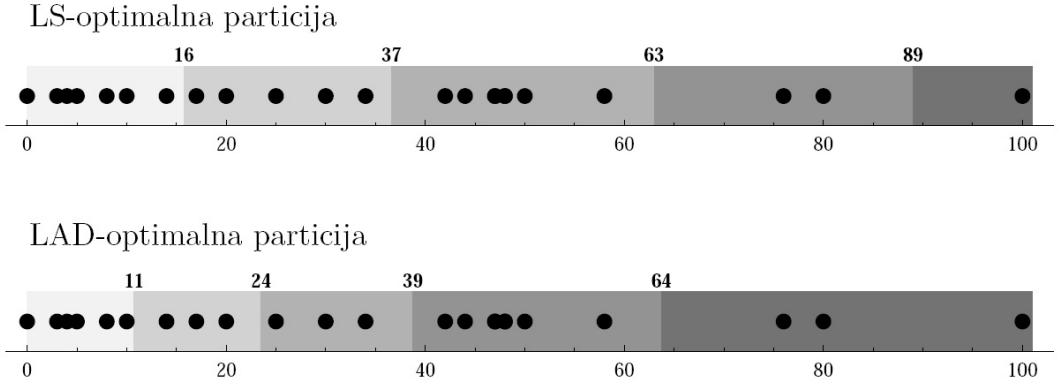
Isto tako, interesantno bi bilo promatrati problem grupiranja učenika temeljem rezultata postignutih na državnoj maturi u okviru jednog predmeta i/ili na osnovi rezultata iz svih predmeta, grupiranje i rangiranje škola na osnovi rezultata državne mature itd.

Primjer 9. *U cilju kontinuiranog praćenja rada studenata u okviru nekog predmeta, tijekom semestra piše se nekoliko kolokvija koji mogu zamijeniti klasični pismeni ispit na kraju semestra. Na osnovi sakupljenih bodova na tim kolokvijima, treba procijeniti prag prolaznosti (ako prije toga nije unaprijed zadan) i*

raspone bodova koji će definirati pozitivne ocjene pismenog dijela ispita: dovoljan (2), dobar (3), vrlo dobar (4), izvrstan (5). Primjerice, neka su zadani sljedeći podaci o postignutim bodovima na kolokvijima za grupu od 23 *studenata* $s_1, \ldots, s_{23}$

$$\mathcal{A} = \{0, 3, 4, 5, 8, 10, 14, 17, 20, 25, 30, 34, 42, 44, 47, 47, 48, 48,\ 50, 58, 76, 80, 100\}.$$

Studente ćemo grupirati u 5 klastera uz primjenu LS i LAD kriterija optimalnosti. Kao što je navedeno u Odjeljku 3.4.2., optimalnu particiju u ovom slučaju dovoljno je tražiti među particijama $\Pi = \{\pi_1, \ldots, \pi_5\}$ čiji klaster se nastavljaju jedan na drugi, tj. među particijama za čije klastere vrijedi: $\max \pi_i < \min \pi_{i+1}$, $i = 1, \ldots, 4$. Ukupni broj takvih particija je 7 315, a kompjutersko pretraživanje svih ovih particija traje svega nekoliko sekundi. Na Slici 4 sivim nijansama označeni su klasteri odgovarajućih LS-optimalnih i LAD-optimalnih particija. Pri tome prvi klaster, označen najsvjetlijom nijansom, sadržava studente koji nisu postigli potrebni prag prolaznosti. Primjećujemo da je LS kriterij optimalnosti sukladno Primjedbi 1 izdvojio izrazito najboljeg studenta u posebni klaster, kojem je pridružena ocjena izvrstan (5), dok je LAD kriterij optimalnosti u klaster najboljih studenata uvrstio čak tri studenta. Sukladno Primjedbi 1, LAD kriterij optimalnosti studente nastoji što ravnomjernije razdijeliti u skupine – klastere.



Slika 4: Optimalne particije bodova postignutih na kolokviju

Primjer 10. *U Tablici 5 prikazane su prosječne ocjene (PO) i odgovarajući brojevi ECTS bodova za* 20 *studenata, koje je potrebno grupirati u pet klastera.*

Student	s_1	s_2	s_3	s_4	s_5	s_6	s_7	s_8	s_9	s_{10}
PO	2.0	2.4	2.7	3.0	3.1	3.1	3.2	3.3	3.5	3.6
ECTS	45	48	47	57	60	55	52	51	50	47
Student	s_{11}	s_{12}	s_{13}	s_{14}	s_{15}	s_{16}	s_{17}	s_{18}	s_{19}	s_{20}
PO	3.7	3.8	3.9	3.9	4.0	4.0	4.3	4.5	4.5	5.0
ECTS	54	49	49	48	46	50	51	50	54	60

Tablica 5: Prosječne ocjene (PO) i ECTS bodovi studenata

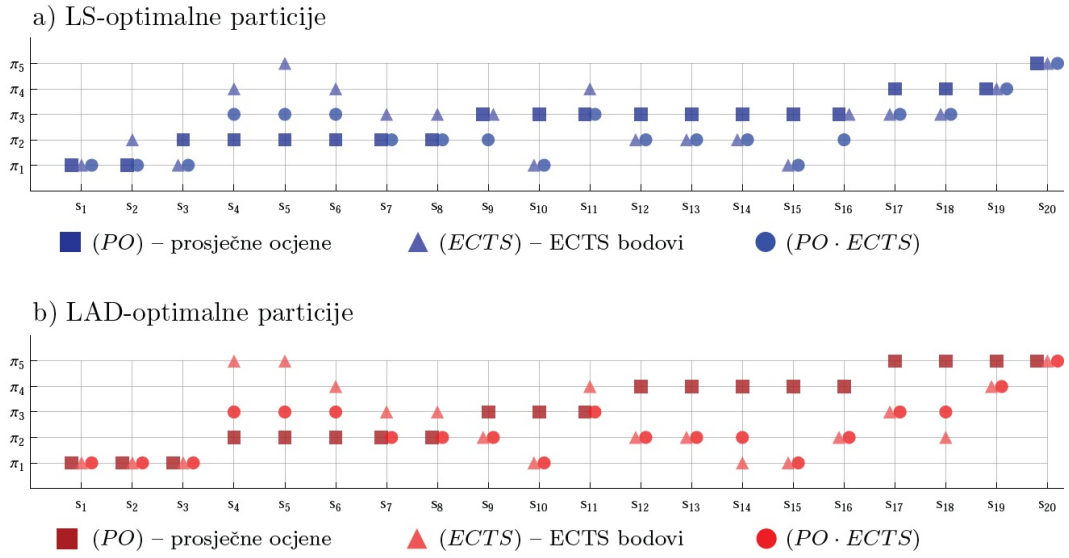
Za rješavanje ovog problema primijenit ćemo metode grupiranja podataka s jednim obilježjem sukladno LS, odnosno LAD kriteriju optimalnosti navedene u Odjeljku 3.. Pri tome studente ćemo grupirati prema:

(i) prosječnim ocjenama;

(ii) postignutom broju ECTS bodova;

(iii) produktu prosječnih ocjena i broja ECTS bodova. Prije toga, podatke o broju ECTS bodova kao i podatke o prosječnim ocjenama potrebno je ravnomjerno normirati kako bi odgovarajući utjecaji na produkt bili podjednaki. Primijetite da su produkti normiranih prosječnih ocjena i normiranih ECTS bodova kvadrati geometrijskih sredina tih podataka.

Kao što je navedeno u Odjeljku 3.4.2., optimalnu particiju u ovom slučaju dovoljno je tražiti među particijama $\Pi = \{\pi_1, \ldots, \pi_5\}$ čiji se klasteri nastavljaju jedan na drugi, tj. među particijama za čije klastere vrijedi: $\max \pi_i < \min \pi_{i+1}$, $i = 1, \ldots, 4$. Ukupni broj takvih particija je 3 876, a kompjutersko pretraživanje svih ovih particija traje svega nekoliko sekundi. Na Slici 5 prikazani su dobiveni rezultati, pri čemu su na osi apscisa naneseni studenti $s_1, \ldots, s_{20}$, dok su na osi ordinata naneseni odgovarajući klasteri $\pi_1, \ldots, \pi_5$. Dobivene LS-optimalne particije označene su plavom bojom (Slika 5a), dok su LAD-optimalne particije prikazane crvenom bojom (Slika 5b). Pri tome su s kvadratićima prikazani rezultati grupiranja studenata prema prosječnim ocjenama, trokutićima su prikazani rezultati grupiranja studenata prema broju ECTS bodova, dok su s krugovima prikazani rezultati grupiranja studenata prema produktu normiranih prosječnih ocjena i broja ECTS bodova.



Slika 5: Grupiranje studenata po različitim kriterijima

Kao što je vidljivo sa Slike 5b, primjenom LAD kriterija optimalnosti studenti s_1, s_2 i s_3, koji imaju izrazito male prosječne ocjene, kao i mali broj ECTS bodova ostaju u klasteru π_1, neovisno o načinu grupiranja (prema prosječnim ocjenama, prema broju ECTS bodova ili prema produktu normiranih prosječnih ocjena i broja ECTS bodova). Slično je sa studentom s_{20}, koji ima najveću moguću prosječnu ocjenu i najveći mogući broj ECTS bodova, te neovisno o načinu grupiranja uvijek ostaje u klasteru π_5. Sukladno Primjedbi 1, LAD kriterij optimalnosti studente nastoji što ravnomjernije razdijeliti u skupine – klastere. Tako primjerice uočavamo da primjenom LAD kriterija optimalnosti kod nekih studenata dolazi do izrazitih oscilacija, u ovisnosti o načinu grupiranja. Primjerice studenti s_4 i s_5, koji imaju visok broj ECTS bodova, a male prosječne ocjene, prilikom grupiranja prema prosječnoj ocjeni smješteni su u klaster π_2, dok su prema broju ECTS bodova smješteni u klaster π_5. Analogna pojava, samo u suprotnom smjeru, prisutna je kod studenata koji imaju visoke prosječne ocjene i malen broj ECTS bodova. Tako primjerice studenti s_{14} i s_{15} iz klastera π_4 prelaze u klaster π_1 dok student s_{18} iz klastera π_5 prelazi u klaster π_2.

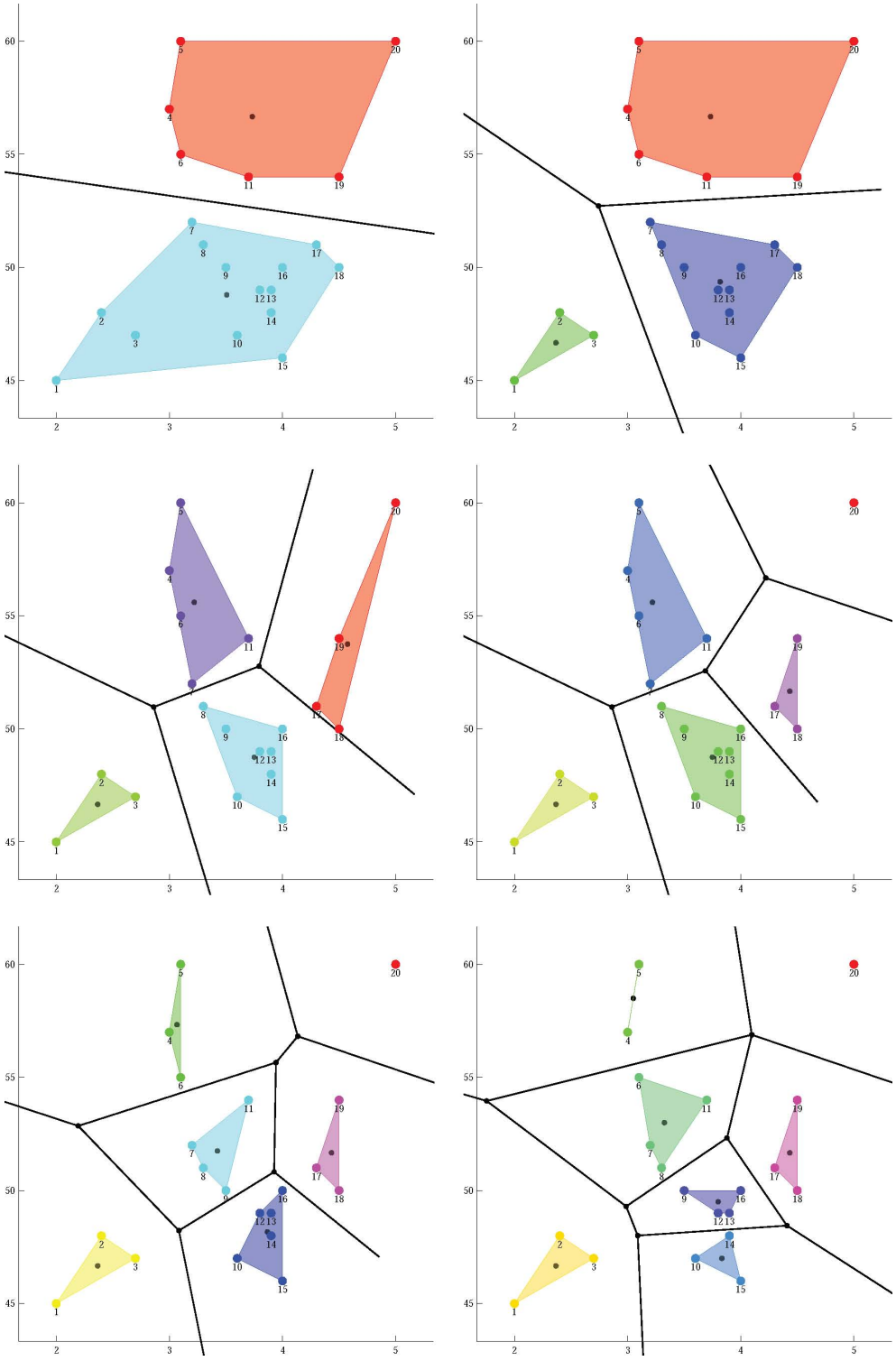
Za razliku od LAD kriterija, primjenom LS kriterija optimalnosti (Slika 5a), značajno je manji broj studenata koji ostaju u istom klasteru. Neovisno o načinu grupiranja, to je student s_1, koji ostaje u klasteru π_1 te student s_{20}, koji ostaje u klasteru π_5. Također, izrazite oscilacije prisutne su samo kod studenta s_5, koji je prilikom grupiranja prema prosječnoj ocjeni smješten u klaster π_2, dok je prema broju ECTS bodova smješten u klaster π_5.

Grupiramo li studente prema produktu normiranih prosječnih ocjena i broja ECTS bodova, rezultat grupiranja prirodno se nalazi između rezultata grupiranja dobivenom prema prosječnim ocjenama i prema broju ECTS bodova, osim u slučaju studenta s_{16} u slučaju LS kriterija optimalnosti (Slika 5a).

Primjer 11. *Studente iz Primjera 10 treba razdijeliti u određen broj klastera koristeći pri tome podatke o prosječnim ocjenama i postignutom broju ECTS bodova iz Tablice 5. Pri tome bi broj klastera trebao biti prirodno određen unutrašnjom strukturom podataka.*

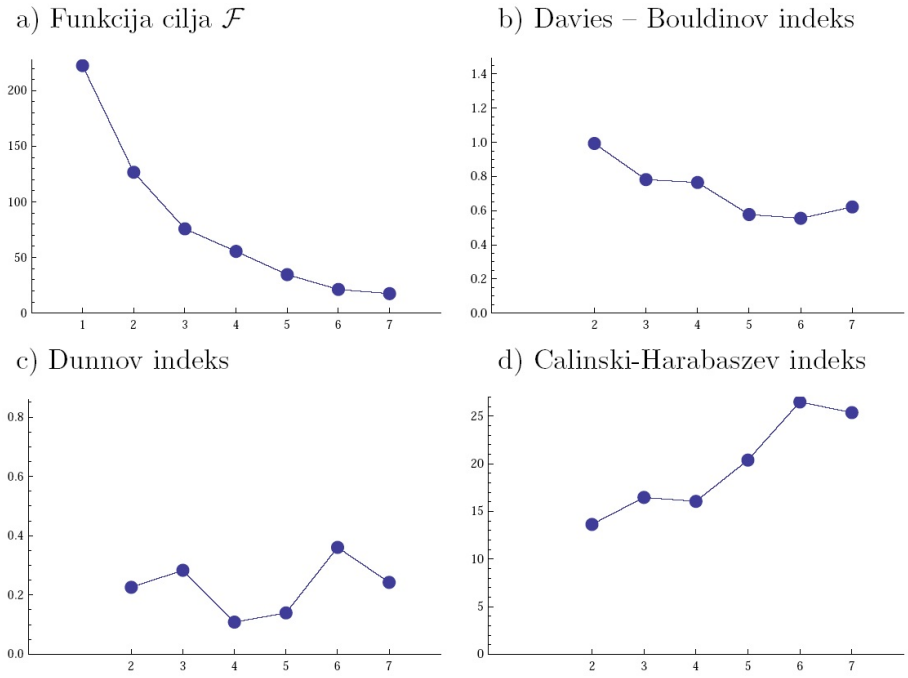
U tu svrhu koristit ćemo samo LS kriterij optimalnosti pri čemu ćemo podatke ravnomjerno normirati (primjerice na segment $[1, 10]$). Na taj način utjecaj prosječne ocjene i broja ECTS bodova bit će ujednačen, a svaki student bit će predstavljen točkom (vektorom) u kvadratu $[1, 10] \times [1, 10] \subset \mathbb{R}^2$. Primjenom LS kriterija optimalnosti opisanog u Odjeljku 4. studente ćemo grupirati redom u $k = 2, \ldots, 7$ klastera.

Ovdje se radi o problemu višedimenzionalne globalne optimizacije nekonveksne funkcije s puno stacionarnih točaka. Nažalost, u ovom slučaju nije realno moguće pretraživati sve particije jer nije moguće značajno reducirati njihov broj kao u slučaju podataka s jednim obilježjem (vidi Odjeljak 3.4.2.(ii)). Optimalne particije potražit ćemo primjenom Algoritma 1, pri čemu ćemo početne particije odrediti na osnovi Slike 5a. Dobivene LS-optimalne particije za $k = 2, \ldots, 7$ prikazane su na Slici 6. Brojevima su označeni studenti po klasterima, a crnim točkicama centroidi klastera.



Slika 6: LS-optimalne particije s $k = 2, \ldots, 7$ klastera

Na kraju, još ćemo pokušati odgovoriti na pitanje koliki broj klastera najbolje odražava unutrašnju strukturu promatranog skupa studenata s prosječnim ocjenama i ECTS bodovima iz Tablice 5. U tu svrhu promatrat ćemo vrijednosti kriterijske funkcije cilja $\mathcal{F}$ i indekse navedene u Odjeljku 5.: Davies – Bouldinov indeks (V_{DB}), Dunnov indeks (V_D) i Calinski-Harabaszev indeks (V_{CH}) za LS-optimalne particije s $2, 3, 4, 5, 6$ i 7 klastera (vidi Sliku 7 i Tablicu 6).



Slika 7: Kretanje indeksa kompaktnosti LS-optimalnih particija s $k = 2, \ldots, 7$ klastera

	Π_2	Π_3	Π_4	Π_5	Π_6	Π_7
$\mathcal{F}$	126.63	75.76	55.50	34.58	21.28	17.51
V_{DB}	0.99	0.78	0.76	0.58	0.55	0.62
V_D	0.23	0.28	0.11	0.14	0.36	0.24
V_{CH}	13.62	16.46	16.05	20.37	26.48	25.36

Tablica 6: Kretanje indeksa kompaktnosti LS-optimalnih particija s $k = 2, \ldots, 7$ klastera

Na osnovi vrijednosti Davies – Bouldinovog (V_{DB}), Dunnovog (V_D) i Calinski-Harabaszovog (V_{CH}) indeksa za LS-optimalne particije s $2, 3, 4, 5, 6$ i 7 klastera vidljivih na Slici 7 i u Tablici 6, može se zaključiti da je LS-optimalna particija sa šest klastera najkompaktnija, a njeni klasteri najbolje separirani i da zbog toga najbolje reprezentira unutrašnju strukturu promatranog skupa studenata s njihovim prosječnim ocjenama i postignutim ECTS bodovima (vidi Sliku 6). To znači da kada bismo htjeli dati LS-optimalne grupne ocjene

promatranih studenata, onda bi to izgledalo ovako:

I. grupa:	s_1, s_2, s_3	– nisu postigli prag prolaznosti
II. grupa:	$s_{10}, s_{12}, s_{13}, s_{14}, s_{15}, s_{16}$	– ocjena E
III. grupa:	s_7, s_8, s_9, s_{11}	– ocjena D
IV. grupa:	s_4, s_5, s_6	– ocjena C
V. grupa:	s_{17}, s_{18}, s_{19}	– ocjena B
VI. grupa:	s_{20}	– ocjena A

7. Programska podrška

Na osnovi izloženog materijala priložena je odgovarajuća programska podrška `Klasteri` izrađena u C++/CLI dostupna na `http://www.mathos.hr/oml/software.htm`. Program `Klasteri` zahtijeva Windows operacijski sustav s .Net Framework 4, a može se snimiti na osobno računalo na sljedeći način:

- S internet adrese `http://www.mathos.hr/oml/software.htm` skinuti arhivu `Klasteri.zip`.
- Raspakirati arhivu i pokrenuti `setup.exe`. Ukoliko .Net Framework 4 nije instaliran, instalacija će ga pokušati skinuti s interneta.
- Odabrati lokaciju na disku gdje želite instalirati program. U odabrani direktorij snimit će se program, primjeri iz ovog teksta i sami tekst u .pdf obliku, a na desktopu pojavit će se shortcut programa.

Programom `Klasteri` moguće je za podatke s jednim obilježjem pronaći LS-optimalne i LAD-optimalne particije s klasterima, njihovim centrima i vrijedosti kriterijske funkcije cilja, a za podatke s dva obilježja lokalno LS-optimalnu particiju s klasterima i njihovim centrima te vrijedosti kriterijske funkcije cilja.

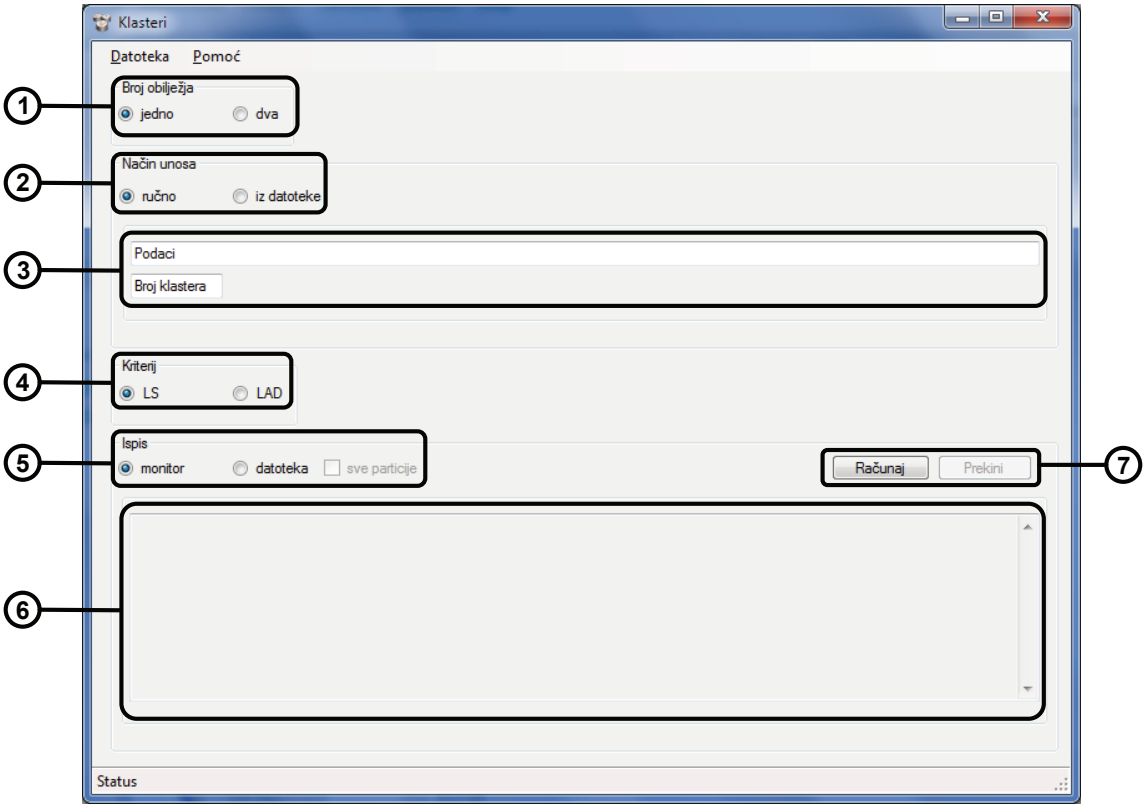
Na Slici 8 označeni su pojedini dijelovi sučelja programa:

1. `[Broj obilježja]` - Odabir vrste podataka prema broju obilježja.
2. `[Način unosa]` - Odabir načina unosa podataka. Podatke je moguće unijeti direktno u sučelje ili učitati iz datoteke.
3. `[Podaci]` - Ovdje se u slučaju odabira ručnog unosa podataka nalaze polja za unos, a u slučaju odabira unosa iz datoteke polje za odabir datoteke s podacima.
4. `[Kriterij]` - Odabir kriterija za računanje kvalitete pojedine particije. Moguće je odabrati kriterij najmanjih kvadrata (LS) ili kriterij najmanjih apsolutnih odstupanja (LAD).
5. `[Ispis]` - Dobiveni rezultati mogu se prikazati na ekranu u polju za ispis ili spremiti u tekstualnu datoteku. U slučaju ispisa u datoteku, može se odabrati ispis svih particija (za podatke s jednim obilježjem), tj. svih koraka k-means algoritma (ako su zadani podaci s dva obilježja).
6. `[polje za ispis]`
7. `[Računaj]` i `[Prekini]` - Stiskom na `[Računaj]` pokreće se postupak traženja optimalnih particija. U slučaju dužeg računanja moguće je stisnuti `[Prekini]` kako bi se program zaustavio.

7.1. Obrada podataka s jednim obilježjem

Podaci se unose na sljedeći način:

- Pod ① odabere se broj obilježja: `[jedno]`;
- Pod ② odabere se željeni način unosa podataka;
- U slučaju odabira ručnog unosa podataka, na ③ pojavit će se dva polja za unos. U prvo polje unose se podaci, odvojeni razmakom. Ukoliko se unose decimalni brojevi, decimalni dio odvaja se točkom. U drugo polje unosi se broj klastera (pozitivan cijeli broj manji ili jednak broju podataka);



Slika 8: Sučelje programa

- Ako je odabran unos podataka iz datoteke, polje za unos putanje do datoteke i dugme za traženje datoteke pojavit će se na ③. Prvi red datoteke mora sadržavati podatke odvojene razmacima, a drugi broj klastera.
- Pod ④ odabere se kriterij, `[LS]` ili `[LAD]`;

Za ispis rezultata mogu se koristiti sljedeće opcije:

- Pod ⑤ moguće je odabrati opciju `[monitor]` ili `[datoteka]`;
- Postupak traženja optimalne particije pokreće se pritiskom na `[Računaj]` (područje ⑦);
- Ako je odabrana opcija `[monitor]`, nakon aktiviranja programa, u polju ⑥ bit će ispisane sve optimalne particije, centri klastera optimalnih particija, vrijednost kriterijske funkcije cilja, broj podataka i broj klastera (primijetimo da u slučaju odabira LAD-kriterija, centri mogu biti brojevi iz nekog segmenata realnih brojeva);
- Ako je odabrana opcija `[datoteka]` i ostavljen je neoznačen kvadratić `[sve particije]`, nakon aktiviranja programa, rezultat koji bi se inače prikazao na monitoru ispisat će se u datoteku;
- Ako je odabrana opcija `[datoteka]` i označen kvadratić `[sve particije]`, nakon aktiviranja programa, u odabranu datoteku bit će ispisane **sve** particije s odgovarajućim centrima klastera i vrijednostima kriterijske funkcije cilja;

- Ako se iz nekog razloga želi prekinuti računanje (predugi postupak u slučaju velikog broja podataka i klastera), pod ⑦ može se stisnuti `[Prekini]`.

Primjer 1. *Zadan je skup podataka* $\mathcal{A} = \{0, 3, 6, 9\}$. *Treba pronaći LS-optimalnu i LAD-optimalnu dvočlanu particiju, centre optimalnih klastera i vrijednost kriterijske funkcije cilja, koristeći program* `Klasteri`.

Prema prethodno napisanim uputama podesimo program za traženje LS-optimalnih particija podataka s jednim obilježjem i unesemo podatke. Program će pronaći jednu LS-optimalnu particiju:

```
INPUT:
0 3 6 9
2
OUTPUT:
Optimalne particije (klasteri i centri):
{{0,3},{6,9}}                                  ⟵ optimalna particija
{1.5, 7.5}                                     ⟵ centri optimalne particije
FLS= 9.0000                                    ⟵ optimalna funkcija cilja
broj podataka:  4  broj particija:  2
```

Za traženje optimalnih particija uz LAD-kriterij, dovoljno je samo pod ⑤ promijeniti kriterij. Program će pronaći tri LAD-optimalne particije:

```
INPUT:
0 3 6 9
2
OUTPUT:
Optimalne particije (klasteri i centri):
{{0},{3,6,9}}                                  ⟵ 1. optimalna particija
{0,6}                                          ⟵ centri 1. optimalne particije
{{0,3},{6,9}}                                  ⟵ 2. optimalna particija
{[0,3],[6,9]}                                  ⟵ centri 2. optimalne particije
{{0,3,6},{9}}                                  ⟵ 3. optimalna particija
{3,9}                                          ⟵ centri 3. optimalne particije
FLAD= 6.0000                                   ⟵ optimalna funkcija cilja
broj podataka:  4  broj particija:  2
```

U svrhu prikaza mogućnosti i ograničenja programa `Klasteri`, za različite vrijednosti broja podataka (m) i broja klastera (k) u Tablici 7 prikazana su vremena izvršavanja programa.

	$m = 50$		$m = 100$		$m = 150$		$m = 200$	
	LS	LAD	LS	LAD	LS	LAD	LS	LAD
$k = 2$	<1s	<1s	<1s	<1s	<1s	<1s	<1s	<1s
$k = 3$	<1s	<1s	<1s	<1s	1.5s	<1s	3s	1.8s
$k = 4$	<1s	<1s	13s	7s	1m5s	32s	3m24s	1m37s
$k = 5$	10s	5.5s	5m16s	2m39s	40m35s	17m	2h48m	1h9m

Tablica 7: Vremena izvršavanja programa za različite brojeve podataka i brojeve klastera na Pentiumu 4, 3.0GHz, 2GB RAM-a

7.2. Obrada podataka s dva obilježja

Podaci se unose na sljedeći način:

- Pod ① odabire se broj obilježja `[dva]`;
- Pod ② odabire se željeni način unosa podataka;
- U slučaju odabira ručnog unosa podataka, na ③ pojavit će se dva polja za unos. U prvo polje unose se podaci. Svaki podatak ima po dva obilježja koja se razdvajaju razmakom, a podaci se razdvajaju vertikalnom crtom (znak |). U drugo polje unose se početni centri u istom obliku kao i podaci.
- Ako je odabran unos podataka iz datoteke, na ③ pojavit će se polje za unos putanje do datoteke i dugme za traženje datoteke. Prvi red datoteke mora sadržavati podatke, a drugi početne centre.
- Pod ④ bit će omogućen samo `[LS]` kriterij;

Za ispis rezultata mogu se koristiti sljedeće opcije:

- Pod ⑤ može se odabrati opcija `[monitor]` ili `[datoteka]`;
- Postupak za traženje lokalno optimalne particije pokreće se pritiskom na `[Računaj]` (područje ⑦);
- Ako je odabrana opcija `[monitor]`, nakon aktiviranja programa, u polju ⑥ ispisat će se lokalno optimalna particija, centri klastera lokalno optimalne particije, vrijednost kriterijske funkcije cilja, broj podataka i broj klastera (primijetite da u nekim slučajevima broj lokalno optimalnih klastera može biti manji od broja početnih centara);
- Ako je odabrana opcija `[datoteka]` i ostavljen neoznačen kvadratić `[svi koraci]`, nakon aktiviranja programa, rezultat koji bi se inače prikazao na monitoru ispisat će se u datoteku;
- Ako je odabrana opcija `[datoteka]` i označen kvadratić `[svi koraci]`, nakon aktiviranja programa, u odabranu datoteku ispisat će se svi koraci k-means algoritma.

Primjer 2. *Zadan je skup podataka* $\mathcal{A} = \{(0,0),(2,1),(1,4),(3,5)\}$. *Krenuvši od particije zadane centrima* $(0,0)$ *i* $(2,1)$ *treba pronaći lokalnu LS-optimalnu dvočlanu particiju, centre njenih klastera i vrijednost kriterijske funkcije cilja, koristeći program* `Klasteri`.

Prema prethodno navedenim uputama podesimo program za traženje LS-optimalnih particija podataka s dva obilježja. Nakon unosa podataka, program će pronaći lokalno LS-optimalnu particiju:

INPUT:
```
0 0 | 2 1 | 1 4 | 3 5
0 0 | 2 1
```
OUTPUT:
```
{{{0,0},{2,1}},{{1,4},{3,5}}}  FLS= 5.0000        ⟵ lok. opt. particija i f. cilja
{{1,0.5},{2,4.5}}                                  ⟵ centri lok. opt. particije
broj podataka:  4  broj particija:  2
```

Ako odaberemo ispis u datoteku i označimo `[svi koraci]`, kao rezultat dobivamo datoteku s niže navedenim sadržajem.

INPUT:
```
0 0 | 2 1 | 1 4 | 3 5
0 0 | 2 1
```
OUTPUT:
```
{{{0,0}},{{2,1},{1,4},{3,5}}} FLS= 10.6667
{{0,0},{2,3.333333}}
{{{0,0},{2,1}},{{1,4},{3,5}}} FLS= 5.0000
{{1,0.5},{2,4.5}}
```

```
{{{0,0},{2,1}},{{1,4},{3,5}}} FLS= 5.0000
{{1,0.5},{2,4.5}}
broj podataka:  4  broj particija:  2
```

Primjedba 7. *Kao što je već ranije navedeno (vidi Odjeljak 3.4.), u slučaju podataka s dva obilježja program* `Klasteri` *neće nužno pronaći optimalnu particiju, već lokalno optimalnu. Pri tome rezultat će bitno ovisiti o odabiru početnih centara.*

Literatura

[1] A. BEN-ISRAEL, C. IYIGUN, *Probabilistic D-clustering*, Journal of Classification **25**(2008), 5–26

[2] D. L. BOYD, L. VANDENBERGHE, *Convex Optimization*, Cambridge University Press, Cambridge, 2004.

[3] T. CALINSKI, J. HARABASZ, *A dendrite method for cluster analysis*, Communications in Statistics, **3**(1974), 1–27

[4] D. DAVIES, D. BOULDIN, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, textbf2(1979), 224–227

[5] I. S. DHILLON, Y. GUAN, B. KULIS, *Kernel k-means, spectral clustering and normalized cuts*, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA, 551–556, 2004

[6] G. DIVÉKI, I. CSANÁD, *Online facility location with facility movements*, CEJOR (2010), DOI 10.1007/s10100-010-0153-8

[7] E. DOMÍNGUEZ, J. MUÑOZ, *Applying bio-inspired techniques to the p-median problem*, IWANN 2005; Computational Intelligence Bioinspired Syst., 8th Int. Workshop Artificial Neural Networks, Lecture Notes in Computer Science, Springer-Verlag, Berlin, 2005, pp. 67 – 74

[8] Z. DREZNER, *Facility Location: A Survey of Aplications and Methods*, Springer-Verlag, Berlin, 2004.

[9] J. DUNN, *Well separated clusters and optimal fuzzy partitions*, Journal of Cybernetics, **4**(1974), 95–104

[10] B. S. EVERITT, S. LANDAU, M. LEESE, *Cluster analysis*, Wiley, London, 2001.

[11] D. E. FINKEL, C. T. KELLEY, *Additive scaling and the* `DIRECT` *algorithm*, J. Glob. Optim. **36**(2006), 597–608

[12] C. A. FLOUDAS, C. E. GOUNARIS, *A review of recent advances in global optimization*, J. Glob. Optim. **45**(2009), 3–38

[13] G. GAN, C MA, J. WU, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.

[14] E. R. HANSEN, G. W. WALSTER, *Global Optimization Using Interval Analysis.* Marcel Dekker, New York, Second Edition, Revised and Expanded, 2004.

[15] M. HUDEC, M. VUJOŠEVIĆ, *A fuzzy system for municipalities classification*, CEJOR **18**(2010), 171–180

[16] C. IYIGUN, A. BEN-ISRAEL, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010), 207–214

[17] C. IYIGUN, *Probabilistic Distance Clustering*, Dissertation, Graduate School – New Brunswick, Rutgers, 2007

[18] D. R. JONES, C. D. PERTTUNEN, B. E. STUCKMAN, *Lipschitzian optimization without the Lipschitz constant*, JOTA **79**(1993), 157–181

[19] J. KOGAN, *Introduction to Clustering Large and High-Dimensional Data*, Cambridge University Press, 2007.

[20] J. KOGAN, C. NICHOLAS, M. WIACEK, *Hybrid Clustering of large high dimensional data*, In M. Castellanos and M. W. Berry (Eds.), Proceedings of the Workshop on Text Mining, SIAM, 2007.

[21] J. KOGAN, M. TEBOULLE, *Scaling clustering algorithms with Bregman distances*. In: M. W. Berry and M. Castellanos (Eds.), Proceedings of the Workshop on Text Mining at the Sixth SIAM International Conference on Data Mining, 2006.

[22] J. KOGAN, C. NICHOLAS, M. WIACEK, *Hybrid clustering with divergences*. In: M. W. Berry and M. Castellanos (Eds.), Survey of Text Mining: Clustering, Classification, and Retrieval, Second Edition, Springer, 2007.

[23] C. IYIGUN, A. BEN-ISRAEL, *A generalized Weiszfeld method for the multi-facility location problem*, Operations Research Letters **38**(2010) 207–214

[24] F LEISCH, *A toolbox for K-centroids cluster analysis*, Computational Statistics & Data Analysis **51**(2006), 526–544

[25] D. LITTAU, D. L. BOLEY, *Clustering very large data sets with PDDP*. In J. Kogan, C. Nicholas, M. Teboulle (eds), *Grouping Multidimensional Data: Recent Advances in Clustering*, 99–126, Springer-Verlag, New York, 2006.

[26] S. PAN,J.,S. CHEN, *Two unconstrained optimization approaches for the Euclidean k-centrum location problem*, Applied Mathematics and Computation **189**(2007) 1368–1383

[27] S. A. PIYAVSKIĬ, *Odin algoritm otyskaniya absolyutnogo ekstrmuma funkcii,* Zh. vychisl. matem. i matem. fiz. **12**(1972), 888-896.

[28] J. REESE, *Solution methods for the p-median problem: an annotated bibliography*, Published online in Wiley InterScience, Wiley, 2006.

[29] A. M. RODRÍGUES-CHIA, I. ESPEJO, Z. DREZNER, *On solving the planar k-centrum problem with Euclidean distances*, EJOR, to appear

[30] K. SABO, R. SCITOVSKI, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal **50**(2008), 185–198

[31] K. SABO, R. SCITOVSKI, I. VAZLER., M. ZEKIĆ-SUŠAC, *Mathematical models of natural gas consumption*, Energy Conversion and Management, (2010), doi.10.1016/jenconman.2010.10.037.

[32] A. SCHÖBEL, D. SCHOLZ, *The big cube small cube solution method for multidimensional facility location problems*, Computers & Operations Research **37**(2010), 115–122

[33] A. SCHÖBEL, *Locating Lines and Hyperplanes: Theory and Algorithms*, Springer Verlag, Berlin, 1999.

[34] B. SHUBERT, *A sequential method seeking the global maximum of a function*, SIAM Journal on Numerical Analysis, **9**(1972), 379–388

[35] H. SPÄTH, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

[36] Z. SU, J. KOGAN, C. NICHOLAS, *Constrained clustering with k-means type algorithms*, In M.W. Berry, J. Kogan (eds), *Text Mining Applications and Theory*, 81–103, Willey, Chichester, 2010.

[37] M. TEBOULLE, *A unified continuous optimization framework for center-based clustering methods*, Journal of Machine Learning Research **8**(2007), 65–102

[38] I. VAZLER, K. SABO, R. SCITOVSKI,, *Weighted median of the data in solving least absolute deviations problems*, Comm. Statist. Theory Methods (to appear)

[39] D. VELJAN, *Kombinatorna i diskretna matematika*, Algoritam, Zagreb, 2001.