

Strojarski fakultet u Slavskom Brodu
Sveučilište J. J. Strossmayera u Osijeku

Matematika i statistika

II. Grupiranje podataka: klasteri

R. Scitovski, M. Benšić, K. Sabo

1 Uvod

Definicija 1. Neka je \mathcal{A} skup s $m \geq 2$ elemenata i $1 \leq k \leq m$. Rastav skupa \mathcal{A} na k podskupove π_1, \dots, π_k , tako da bude

- (i) $\bigcup_{i=1}^k \pi_i = \mathcal{A}$,
- (ii) $\pi_i \cap \pi_j = \emptyset, \quad i \neq j$,
- (iii) $m_j := |\pi_j| \geq 1, \quad j = 1, \dots, k$.

zovemo *particija skupa \mathcal{A}* , a skupove π_1, \dots, π_k *klasteri*. Skup svih particija skupa \mathcal{A} sastavljenih od k klastera koje zadovoljavaju (i)-(iii) 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 ovakvih podskupova skupa \mathcal{A} . Na taj način svjesno smo iz razmatranja isključili particije, koje sadržavaju prazan skup ili skup \mathcal{A} .

Sinonimi: *grupiranje, segmentiranje, klasifikacija, rangiranje*

En.: cluster analysis, clustering, data mining

Može se pokazati da je broj svih particija skupa \mathcal{A} iz Definicije 1 jednak Stirlingovom broju druge vrste

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

Primjer 1. Broj svih particija skupa \mathcal{A} koje zadovoljavaju Definiciju 1 specijalno za $m = 10, 50, 10^3, 10^6$ i $k = 2, 3, 5, 8, 10$ iznosi

$ \mathcal{P}(\mathcal{A}, k) $	$k = 2$	$k = 3$	$k = 5$	$k = 8$	$k = 10$
$m = 10$	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 teba reći da problem traženja optimalne particije spada u NP-teške probleme [Gan et al., 2007] nekonveksne optimizacije općenito nediferencijabilne funkcije više varijabli, koja najčešće posjeduje značajan broj stacionarnih točaka.

Primjene:

poljoprivreda (primjerice, razvrstavanje oranica prema plodnosti zemljišta);

biologija (primjerice, klasifikacija kukaca u grupe)

medicina (primjerice, analiza rendgenskih slika)

promet (primjerice, identifikacija prometnih “čepova”)

analiza i pretraživanje teksta

analiza klimatskih kretanja

donošenje raznih odluka u tijelima državne i lokalne administracije.

definiranje izbornih sustava

Programska podrška [Sabo et al., 2010]:

<http://www.mathos.hr/oml/software.htm>

2 Motivacija

$\mathcal{A} = \{a_1, \dots, a_m\} \subset \mathbb{R}$ – podskup realnih brojeva

$\Pi(\mathcal{A}) = \{\pi_1, \pi_2\}$ – particijaskupa \mathcal{A} , takva da vrijedi

$$\pi_1 \cup \pi_2 = \mathcal{A}, \quad \pi_1 \cap \pi_2 = \emptyset, \quad m_1 = |\pi_1| \geq 1, \quad m_2 = |\pi_2| \geq 1.$$

$|\mathcal{P}(\mathcal{A}, k)| = 2^{m-1} - 1$ – broj svih ovakvih particija

$d: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$ – kvazimetrička funkcija

Uz primjenu LS-kvazimetričke funkcije $d(x, y) = (x - y)^2$, za klastere π_1, π_2 odredimo reprezentante – centroide

$$c_1 = \operatorname{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_1} d(x, a_i) = \frac{1}{m_1} \sum_{a_i \in \pi_1} a_i,$$

$$c_2 = \operatorname{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_2} d(x, a_i) = \frac{1}{m_2} \sum_{a_i \in \pi_2} a_i.$$

Mjeru kvalitete particije $\Pi = \{\pi_1, \pi_2\}$ s centroidima c_1, c_2 definirat ćemo funkcijom cilja $\mathcal{F}: \mathcal{P}(\mathcal{A}, 2) \rightarrow \mathbb{R}_+$

$$\mathcal{F}(\Pi) = \sum_{a_i \in \pi_1} d(c_1, a_i) + \sum_{a_i \in \pi_2} d(c_2, a_i), \quad (2)$$

koja predstavlja sumu “kvadratnog rasipanja” točaka klastera π_1 do centroida c_1 i točaka klastera π_2 do centroida c_2 .

Problem: *pronaći onu particiju na kojoj funkcija cilja \mathcal{F} postiže najmanju vrijednost*

Primjer 2. *Treba pronaći sve particije skupa $\mathcal{A} = \{1, 3, 4, 8\}$, odrediti pripadne centroide i vrijednosti funkcije cilja \mathcal{F} .*

π_1	π_2	c_1	c_2	$\mathcal{F}(\Pi)$		$\mathcal{G}(\Pi)$	
{1}	{3, 4, 8}	1	5	0 + 14 =	14	9 + 3 =	12
{3}	{1, 4, 8}	3	13/3	0 + 74/3 =	24.67	1 + 1/3 =	1.33
{4}	{1, 3, 8}	4	4	0 + 26 =	26	0 + 0 =	0
{8}	{1, 3, 4}	8	8/3	0 + 14/3 =	4.67	16 + 16/3 =	21.33
{1, 3}	{4, 8}	2	6	2 + 8 =	10	8 + 8 =	16
{1, 4}	{3, 8}	5/2	11/2	9/2 + 25/2 =	17	9/2 + 9/2 =	9
{1, 8}	{3, 4}	9/2	7/2	49/2 + 1/2 =	25	1/2 + 1/2 =	1

Obratno, za dane realne brojeve $c_1, c_2 \in \mathbb{R}$, $c_1 \neq c_2$, primjenom **principa minimalnih udaljenosti** možemo definirati particiju $\Pi = \{\pi_1, \pi_2\}$ skupa \mathcal{A} na sljedeći način:

$$\pi_1 = \{a_i \in \mathcal{A} : d(a_i, c_1) \leq d(a_i, c_2)\},$$

$$\pi_2 = \{a_i \in \mathcal{A} : d(a_i, c_2) < d(a_i, c_1)\},$$

pri čemu treba voditi računa da svaki element skupa \mathcal{A} pridružimo samo jednom klasteru. Zato se problem traženja optimalne particije skupa \mathcal{A} može razmatrati kao sljedeći optimizacijski problem

$$\min_{c_1, c_2 \in \mathbb{R}} F(c_1, c_2), \quad F(c_1, c_2) = \sum_{i=1}^m \min\{d(c_1, a_i), d(c_2, a_i)\}, \quad (3)$$

gdje je $F: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$. Ovaj problem u literaturi se pojavljuje pod nazivom *k-median problem* i ekvivalentan je problemu traženja optimalne particije na kojoj kriterijska funkcija cilja \mathcal{F} postiže globalni minimum.

Primjedba 1. Primijetite da klasteri π_1, π_2 ovise o centrima c_1, c_2 i da vrijedi

$$\mathcal{F}(\Pi) := \sum_{a_i \in \pi_1} d(c_1, a_i) + \sum_{a_i \in \pi_2} d(c_2, a_i) \leq \sum_{i=1}^m \min\{d(c_1, a_i), d(c_2, a_i)\} =: F(c_1, c_2), \quad (4)$$

pri čemu se jednakost postiže na lokalno optimalnoj particiji. Naime, vrijedi

$$\begin{aligned} F(c_1, c_2) &= \sum_{a_i \in \pi_1(c_1, c_2)} \min\{d(c_1, a_i), d(c_2, a_i)\} + \sum_{a_i \in \pi_2(c_1, c_2)} \min\{d(c_1, a_i), d(c_2, a_i)\} \\ &\geq \sum_{a_i \in \pi_1(c_1, c_2)} d(c_1, a_i) + \sum_{a_i \in \pi_2(c_1, c_2)} d(c_2, a_i) = \mathcal{F}(\Pi). \end{aligned}$$

Algoritam 1. (Standardni k-means algorithm)Step 1: Inicijalizacija: $z_1 < z_2$;

$$F(z_1, z_2) = \sum_{i=1}^m \min\{d(z_1, a_i), d(z_2, a_i)\}$$

Step 2: Priduživanje (assignment step)

$$\pi_1 = \{a_i \in \mathcal{A} : d(z_1, a_i) \leq d(z_2, a_i)\},$$

$$\pi_2 = \{a_i \in \mathcal{A} : d(z_2, a_i) < d(z_1, a_i)\},$$

Step 3: Korekcija (update step)

$$\zeta_1 = \operatorname{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_1} d(x, a_i)$$

$$\zeta_2 = \operatorname{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_2} d(x, a_i)$$

$$\mathcal{F}(\Pi(\pi_1, \pi_2)) = \sum_{a \in \pi_1} d(\zeta_1, a) + \sum_{a \in \pi_2} d(\zeta_2, a)$$

Primjedba 2. Step 2 i Step 3 se izmjenjuju tako dugo dok se ili centri ne poklope ili dok se particije ne poklope ili dok vrijednost funkcije cilja ne prestane opadati. Primijetite da se u Step 2 novi klasteri π_1, π_2 geometrijski mogu odrediti tako da odredimo polovište spojnice centroida. Tada svi elementi lijevo od polovišta pripadaju novom klasteru π_1 , a svi elementi desno od polovišta pripadaju novom klasteru π_2 .

Primjer 3. $\mathcal{A} = \{1, 2, 6, 7, 9\}$, $k = 2$,

R.br.	z_1	z_2	$F(z_1, z_2)$	π_1	π_2	ζ_1	ζ_2	$\mathcal{F}(\{\pi_1, \pi_2\})$
1	4	8	19	$\{1, 2, 6\}$	$\{7, 9\}$	3	8	16
2	3	8	11	$\{1, 2\}$	$\{6, 7, 9\}$	$\frac{3}{2}$	$\frac{22}{3}$	$\frac{31}{6} \approx 5.1667$
3	$\frac{3}{2}$	$\frac{22}{3}$	5.1667	$\{1, 2\}$	$\{6, 7, 9\}$	$\frac{3}{2}$	$\frac{22}{3}$	5.1667

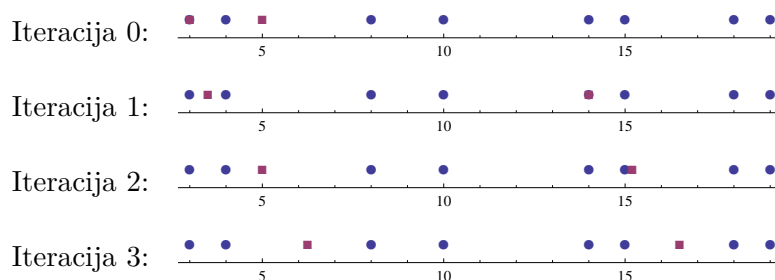
Primjer 4. $\mathcal{A} = \{0, 2, 3\}$, $k = 2$,

R.br.	z_1	z_2	$F(z_1, z_2)$	π_1	π_2	ζ_1	ζ_2	$\mathcal{F}(\{\pi_1, \pi_2\})$
1	1	3	2	$\{0, 2\}$	$\{3\}$	1	3	2
1	1	3	2	$\{0, 2\}$	$\{3\}$	1	3	2

Ovo je primjer koji pokazuje da standardni k -means algoritam ne daje optimalno rješenje. Bolja particija je $\Pi_1 = \{\{0\}, \{2, 3\}\}$ jer je $F(\Pi_1) = 0.25$.

Primjer 5. $\mathcal{A} = \{3, 4, 8, 10, 14, 15, 18, 19\}$, $k = 2$,

R.br.	z_1	z_2	$F(z_1, z_2)$	π_1	π_2	ζ_1	ζ_2	$\mathcal{F}(\{\pi_1, \pi_2\})$
1	3	5	581	{3, 4}	{8, 10, 14, 15, 18, 19}	3.5	14	94.5
2	3.5	14	78.75	{3, 4, 8}	{10, 14, 15, 18, 19}	5	15.2	64.8
3	5	15.2	62.76	{3, 4, 8, 10}	{14, 15, 18, 19}	6.25	16.5	49.75
4	6.25	16.5	49.756	{3, 4, 8, 10}	{14, 15, 18, 19}	6.25	16.5	49.75



2.1 Traženje optimalne particije

Zadatak 1. Neka su $f, g \in C^2(\mathbb{R})$ dvije funkcije za koje vrijedi

$$f(x) + g(x) = \text{const}, \quad \forall x \in \mathbb{R}.$$

Pokažite da tada vrijedi

- (i) Ako funkcija f u točki $x_0 \in \mathbb{R}$ postiže lokalni (odnosno globalni) minimum, onda funkcija g u toj točki postiže lokalni (odnosno globalni) maksimum;
- (ii) Ako funkcija f u točki $x_0 \in \mathbb{R}$ postiže lokalni (odnosno globalni) maksimum, onda funkcija g u toj točki postiže lokalni (odnosno globalni) minimum;

Lema 1. Neka je $\mathcal{A} = \{a_1, \dots, a_m\} \subset \mathbb{R}$ skup podataka, a $\Pi = \{\pi_1, \pi_2\}$ neka particija s klasterima π_1, π_2 i centrima

$$c_1 = \frac{1}{m_1} \sum_{a_i \in \pi_1} a_i, \quad c_2 = \frac{1}{m_2} \sum_{a_i \in \pi_2} a_i, \quad m_1 = |\pi_1|, \quad m_2 = |\pi_2|.$$

Tada vrijedi

$$\mathcal{F}(c_1, c_2) + \mathcal{G}(c_1, c_2) = \sum_{i=1}^m (a_i - \bar{c})^2, \quad \bar{c} = \frac{1}{m} \sum_{i=1}^m a_i, \quad (5)$$

gdje je

$$\mathcal{F}(c_1, c_2) = \sum_{a_i \in \pi_1} (a_i - c_1)^2 + \sum_{a_i \in \pi_2} (a_i - c_2)^2, \quad (6)$$

$$\mathcal{G}(c_1, c_2) = m_1(c_1 - \bar{c})^2 + m_2(c_2 - \bar{c})^2. \quad (7)$$

Dokaz. Za svaki $x \in \mathbb{R}$ vrijedi

$$\sum_{a_i \in \pi_j} (a_i - x)^2 = \sum_{a_i \in \pi_j} (a_i - c_j)^2 + m_j (c_j - x)^2, \quad j = 1, 2. \quad (8)$$

Primjerice,

$$\begin{aligned} \sum_{a_i \in \pi_1} (a_i - x)^2 &= \sum_{a_i \in \pi_1} ((a_i - c_1) + (c_1 - x))^2 \\ &= \sum_{a_i \in \pi_1} (a_i - c_1)^2 + m_1 (c_1 - x)^2, \end{aligned}$$

jer je

$$\sum_{a_i \in \pi_1} (a_i - c_1)(c_1 - x) = (c_1 - x) \sum_{a_i \in \pi_1} (a_i - c_1) = 0.$$

Ako u (8) stavimo $x = \bar{c}$ i zbrojimo obje jednakosti, dobivamo

$$\sum_{a_i \in \pi_1} (a_i - \bar{c})^2 + \sum_{a_i \in \pi_2} (a_i - \bar{c})^2 = \sum_{a_i \in \pi_1} (a_i - c_1)^2 + \sum_{a_i \in \pi_2} (a_i - c_2)^2 + m_1 (c_1 - \bar{c})^2 + m_2 (c_2 - \bar{c})^2.$$

Kako je

$$\sum_{a_i \in \pi_1} (a_i - \bar{c})^2 + \sum_{a_i \in \pi_2} (a_i - \bar{c})^2 = \sum_{i=1}^m (a_i - \bar{c})^2$$

slijedi (5). □

Teorem 1. *Neka su $F, G \in C^2(\mathbb{R}^2)$ dvije funkcije za koje vrijedi*

$$F(x, y) + G(x, y) = \text{const}, \quad \forall (x, y) \in \mathbb{R}^2.$$

Tada vrijedi:

- (i) *Ako funkcija F u točki $(x^*, y^*) \in \mathbb{R}^2$ postiže lokalni (odnosno globalni) minimum, onda funkcija G u toj točki postiže lokalni (odnosno globalni) maksimum;*
- (ii) *Ako funkcija F u točki $(x^*, y^*) \in \mathbb{R}^2$ postiže lokalni (odnosno globalni) maksimum, onda funkcija G u toj točki postiže lokalni (odnosno globalni) minimum;*

Dokaz. Kako je

$$\frac{\partial F}{\partial x} + \frac{\partial G}{\partial x} = 0, \quad \frac{\partial F}{\partial y} + \frac{\partial G}{\partial y} = 0,$$

onda se stacionarne točke funkcija F i G podudaraju. Nadalje, kako je

$$\frac{\partial^2 F}{\partial x^2} + \frac{\partial^2 G}{\partial x^2} = 0, \quad \frac{\partial^2 F}{\partial y \partial x} + \frac{\partial^2 G}{\partial x \partial y} = 0, \quad \frac{\partial^2 F}{\partial y^2} + \frac{\partial^2 G}{\partial y^2} = 0,$$

onda za Hessijane funkcija F, G vrijedi

$$H_F(x, y) = -H_G(x, y),$$

iz čega slijede traženi zaključci. □

Zadatak 2. Neka su $F, G \in C(\mathbb{R})$ dvije neprekidne funkcije za koje vrijedi

$$F(x, y) + G(x, y) = \text{const}, \quad \forall (x, y) \in \mathbb{R}^2.$$

Pokažite da tada vrijedi:

(i) Ako funkcija F u točki $(x^*, y^*) \in \mathbb{R}^2$ postiže lokalni (odnosno globalni) minimum, onda funkcija G u toj točki postiže lokalni (odnosno globalni) maksimum;

(ii) Ako funkcija F u točki $(x^*, y^*) \in \mathbb{R}^2$ postiže lokalni (odnosno globalni) maksimum, onda funkcija G u toj točki postiže lokalni (odnosno globalni) minimum;

Korolar 1. Neka je $\mathcal{A} = \{a_1, \dots, a_m\} \subset \mathbb{R}$ skup podataka. Tada optimalnu dvočlanu particiju $\Pi^* = \{\pi_1^*, \pi_2^*\}$ skupa \mathcal{A} možemo tražiti tako da tražimo globalni minimum funkcije (princip minimalnih kvadratnih udaljenosti od centara klastera)

$$\mathcal{F}(c_1, c_2) = \sum_{a_i \in \pi_1} (a_i - c_1)^2 + \sum_{a_i \in \pi_2} (a_i - c_2)^2,$$

gdje je

$$\pi_1 = \{a_i \in \mathcal{A} : |a_i - c_1| \leq |a_i - c_2|\}, \quad \pi_2 = \{a_i \in \mathcal{A} : |a_i - c_2| < |a_i - c_1|\},$$

odnosno, sukladno Primjedbi 1, minimum funkcije

$$F(c_1, c_2) = \sum_{i=1}^m \min\{(a_i - c_1)^2, (a_i - c_2)^2\},$$

ili tako da tražimo globalni maksimum funkcije (princip maksimalne kvadratne udaljenosti centara klastera)

$$\mathcal{G}(c_1, c_2) = m_1(c_1 - \bar{c})^2 + m_2(c_2 - \bar{c})^2, \quad \bar{c} = \frac{1}{m} \sum_{i=1}^m a_i,$$

gdje je $m_1 = |\pi_1|$, $m_2 = |\pi_2|$.

Primjedba 3. Fizikalni smisao bio bi traženje “dva težišta” diskretnog materijalnog tijela.

3 Grupiranje podataka s jednim obilježjem

Neka je $\mathcal{A} = \{a_1, \dots, a_m\}$ skup koji na osnovi samo jednog obilježja treba grupirati u k klastera koji zadovoljavaju Definiciju 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 $\mathcal{A} = \{a_1, \dots, a_m\}$ među kojima može biti i jednakih. Možemo također koristiti i termine: *m-torka realnih brojeva* ili *konačni niz realnih brojeva*.

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

$$c_j = c(\pi_j) := \operatorname{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_j} d(x, a_i). \quad (9)$$

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

$$\mathcal{F}(\Pi) = \sum_{j=1}^k \sum_{a_i \in \pi_j} d(c_j, a_i), \quad (10)$$

onda d -optimalnu particiju Π^* tražimo rješavanjem sljedećeg optimizacijskog problema

$$\mathcal{F}(\Pi^*) = \min_{\Pi \in \mathcal{P}(\mathcal{A}, k)} \mathcal{F}(\Pi). \quad (11)$$

Primijetite da na taj način optimalna particija Π^* 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.

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

$$\pi_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a), \forall s = 1, \dots, k\}, \quad j = 1, \dots, k, \quad (12)$$

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

$$\min_{c_1, \dots, c_k \in \mathbb{R}} F(c_1, \dots, c_k), \quad F(c_1, \dots, c_k) = \sum_{i=1}^m \min_{j=1, \dots, k} d(c_j, a_i), \quad (13)$$

gdje je $F: \mathbb{R}^k \rightarrow \mathbb{R}_+$. Općenito, ova funkcija nije konveksna ni diferencijabilna, a može imati više lokalnih minimuma [Gan et al., 2007, Iyigun and Ben-Israel, 2010, Teboulle, 2007].

Optimizacijski problem (13) u literaturi se može naći pod nazivom *k-median problem* i ekvivalentan je optimizacijskom problemu (11). Naime, vrijedi

$$\begin{aligned} F(c_1, \dots, c_k) &:= \sum_{i=1}^m \min\{d(c_1, a_i), \dots, d(c_k, a_i)\} \\ &\geq \sum_{j=1}^k \sum_{a_i \in \pi_j} \min\{d(c_1, a_i), \dots, d(c_k, a_i)\} \\ &= \sum_{j=1}^k \sum_{a_i \in \pi_j} d(c_j, a_i) =: \mathcal{F}(\Pi), \end{aligned} \quad (14)$$

gdje je $\Pi = \{\pi_1, \dots, \pi_k\}$,

$$\pi_j = \pi_j(c_1, \dots, c_k) = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a), \forall s = 1, \dots, k\},$$

odnosno

$$\pi_j = \pi_j(c_1, \dots, c_k) = \{a_i \in \mathcal{A} : j = \operatorname{argmin}_{s=1, \dots, k} d(c_s, a_i)\}.$$

Jednakost u (14) vrijedi ako je Π lokalno ili globalno optimalna particija.

3.1 Kriterij najmanjih kvadrata

Ako je $d: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$, $d(a, b) = (a - b)^2$ LS-kvazimetrička funkcija, centri c_1, \dots, c_k klastera π_1, \dots, π_k određeni su s

$$c_j = \operatorname{argmin}_{u \in \mathbb{R}} \sum_{a_i \in \pi_j} (a_i - u)^2 = \frac{1}{|\pi_j|} \sum_{a_i \in \pi_j} a_i, \quad j = 1, \dots, k, \quad (15)$$

a funkcija cilja (10) s

$$\mathcal{F}(c_1, \dots, c_k) = \sum_{j=1}^k \sum_{a_i \in \pi_j} (c_j - a_i)^2. \quad (16)$$

Primjer 6. Zadan je skup $\mathcal{A} = \{0, 3, 6, 9\}$. Treba pronaći sve njegove dvočlane particije koje zadovoljavaju Definiciju 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 2 LS-optimalna particija u ovom slučaju je $\{\{0, 3\}, \{6, 9\}\}$ jer na njoj kriterijska funkcija cilja \mathcal{F} zadana s (16) postiže globalni minimum.

3.1.1 Dualni problem

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

π_1	π_2	c_1	c_2	$\mathcal{F}(c_1, c_2)$		$\mathcal{G}(c_1, c_2)$	
{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	= 1

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

Lema 2. Neka je $\mathcal{A} = \{a_1, \dots, a_m\}$ skup podataka, a $\Pi = \{\pi_1, \dots, \pi_k\}$ neka particija s klasterima π_1, \dots, π_k duljine m_1, \dots, m_k . Neka je nadalje

$$c = \frac{1}{m} \sum_{i=1}^m a_i, \quad c_j = \frac{1}{m_j} \sum_{a_i \in \pi_j} a_i, \quad j = 1, \dots, k, \quad (17)$$

gdje je $m_j = |\pi_j|$. Tada vrijedi

$$\sum_{i=1}^m (a_i - c)^2 = \mathcal{F}(c_1, \dots, c_k) + \mathcal{G}(c_1, \dots, c_k), \quad (18)$$

gdje je

$$\mathcal{F}(c_1, \dots, c_k) = \sum_{j=1}^k \sum_{a_i \in \pi_j} (c_j - a_i)^2, \quad (19)$$

$$\mathcal{G}(c_1, \dots, c_k) = \sum_{j=1}^k m_j (c_1, \dots, c_k) (c_j - c)^2. \quad (20)$$

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

$$\sum_{a_i \in \pi_j} (a_i - x)^2 = \sum_{a_i \in \pi_j} (a_i - c_j)^2 + m_j (c_j - x)^2, \quad j = 1, \dots, k. \quad (21)$$

Naime, kako je $\sum_{a_i \in \pi_j} (a_i - c_j)(c_j - x) = (c_j - x) \sum_{a_i \in \pi_j} (a_i - c_j) = 0$, 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 (21) umjesto x stavimo $c = \frac{1}{m} \sum_{i=1}^m a_i$ i zbrojimo sve jednakosti, dobivamo (18). \square

Iz Leme 2 neposredno slijedi tvrdnja sljedećeg teorema [Dhillon et al., 2004, Späth, 1983]

Teorem 2. Uz oznake kao u Lemi 2 vrijedi:

$$\begin{aligned} \operatorname{argmin}_{\Pi \in \mathcal{P}(\mathcal{A}, k)} \mathcal{F}(\Pi) &= \operatorname{argmax}_{\Pi \in \mathcal{P}(\mathcal{A}, k)} \mathcal{G}(\Pi), \\ \operatorname{argmin}_{c_1, \dots, c_k \in \mathbb{R}} \mathcal{F}(c_1, \dots, c_k) &= \operatorname{argmax}_{c_1, \dots, c_k \in \mathbb{R}} \mathcal{G}(c_1, \dots, c_k). \end{aligned}$$

To znači da u cilju pronalaženja LS-optimalne particije, umjesto minimizacije funkcije \mathcal{F} zadane s (16), odnosno (19), možemo maksimizirati funkciju

$$\mathcal{G}(c_1, \dots, c_k) = \sum_{j=1}^k m_j (c_1, \dots, c_k) (c_j - c)^2. \quad (22)$$

Primjer 7. Skup $A = \{0, 3, 6, 9\}$ iz Primjera 6 ima 7 različitih particija i za sve njih u Tablici 2 prikazana je vrijednost kriterijske funkcije cilja \mathcal{G} . Kao što se vidi iz Tablice 2 funkcija \mathcal{G} prima maksimalnu vrijednost na optimalnoj particiji $\{\{0, 3\}, \{6, 9\}\}$, što je u skladu s Teoremom 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 (17) dana s

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

Specijalno, za dva disjunktna skupa realnih brojeva $A = \{x_1, \dots, x_p\}$, $B = \{y_1, \dots, y_q\}$ 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

Ako je $d: \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}_+$, $d(a, b) = |a - b|$ l_1 -metrička funkcija definirana, centri c_1, \dots, c_k klastera π_1, \dots, π_k određeni su s

$$c_j = \operatorname{argmin}_{u \in \mathbb{R}} \sum_{a_i \in \pi_j} |a_i - u| = \operatorname{med}(\pi_j), \quad j = 1, \dots, k, \quad (23)$$

a funkcija cilja (10), odnosno (13) s

$$\mathcal{F}(c_1, \dots, c_k) = \sum_{j=1}^k \sum_{a_i \in \pi_j} |c_j - a_i|. \quad (24)$$

Ako pri tome iskoristimo (25), onda za izračunavanje funkcije cilja (24) nije potrebno poznavati centre klastera (23), što može značajno ubrzati računski proces.

Zadatak 3. Neka je $\mathcal{A} = \{a_1, \dots, a_m\}$ konačan niz realnih brojeva. Pokažite da vrijedi

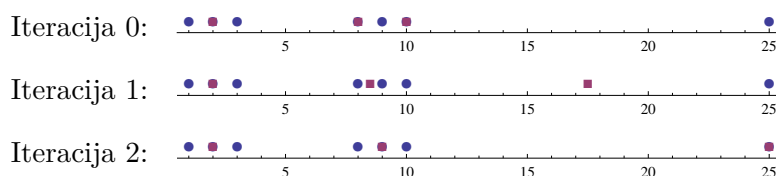
$$\sum_{i=1}^m |a_i - \operatorname{med}(A)| = \sum_{i=1}^k (a_{m-i+1} - a_i). \quad (25)$$

Primjedba 5. U slučaju izbora LAD-kriterija optimalnosti u Koraku 2 Algoritma 2 može se dogoditi da neki centroid ζ_j , može biti proizvoljan broj iz nekog intervala $[\alpha, \beta] \subset \mathbb{R}$. U tom slučaju treba uzeti $\zeta_j = \frac{\alpha + \beta}{2}$.

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

R.br.	z_1	z_2	z_3	$F(z_1, z_2, z_3)$	π_1	π_2	π_3	ζ_1	ζ_2	ζ_3	$\mathcal{F}(\{\pi_1, \pi_2, \pi_3\})$
1	2	8	10	228	{1, 2, 3}	{8, 9}	{10, 25}	2	8.5	17.5	115
2	2	8.5	17.5	61	{1, 2, 3}	{8, 9, 10}	{25}	2	9	25	4
3	2	9	25	4	{1, 2, 3}	{8, 9, 10}	{25}	2	9	25	4

Broj svih tročlanih particija ovog skupa je $\frac{1}{2}(3^m - 2^m + 1) = 301$. U prethodnoj tablici prikazan je tijek iterativnog postupka. Direktnom provjerom svih particija može se pokazati da je Algoritam 2 pronašao upravo optimalnu particiju. Iz ovog primjera vidljivo je da k -means algoritam u smislu LS-optimalnosti daje particiju, koja značajno ovisi o stršećem podatku, tako da upravo stršeći podatak čini zaseban klaster (vidi Sliku 1).



Slika 1: k -means algoritam za traženje lokalno optimalne particije skupa \mathcal{A} iz Primjera 8

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

R.br.	z_1	z_2	$F(z_1, z_2)$	π_1	π_2	ζ_1	ζ_2	$\mathcal{F}(\{\pi_1, \pi_2\})$
1	2	15	29	{1, 2, 3, 8}	{9, 10, 25}	[2, 3]	10	24
2	[2, 3]	10	20	{1, 2, 3}	{8, 9, 10, 25}	2	[9, 10]	20
3	2	[9, 10]	20	{1, 2, 3}	{8, 9, 10, 25}	2	[9, 10]	20

Primjenom Algoritma 2 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 je prikazan tijek iterativnog postupka. Pri tome, centri u Koraku 2 Algoritma 2 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 prirodno je pridružen drugom klasteru (vidi Sliku 2).

3.3 Grupiranje podataka s težinama

Pretpostavimo da je zadan skup podataka $\mathcal{A} = \{a_1, \dots, a_m\}$, pri čemu je svakom podatku a_i pridružena odgovarajuća težina $w_i > 0$. Kriterijska funkcija cilja (10) sada postaje

$$\mathcal{F}(\Pi) = \sum_{j=1}^k \sum_{a_i \in \pi_j} w_i d(c_j, a_i). \quad (26)$$

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

$$c_j = \frac{1}{\kappa_j} \sum_{a_i \in \pi_j} w_i a_i, \quad \kappa_j = \sum_{a_i \in \pi_j} w_i, \quad (27)$$

a kod primjene kriterija LAD-optimalnosti centar c_j klastera π_j određen je težinskim medijanom podataka koji pripadaju klasteru π_j [Sabo and Scitovski, 2008, Vazler et al., 2012]

$$c_j = \text{med}_{a_i \in \pi_j}(w_i, a_i). \quad (28)$$

Algoritam 2. (Standardni k-means algoritam)

Step 1: Učitati m , k , elemente skupa \mathcal{A} i težine $w_1, \dots, w_m > 0$;

Izabrati: $\min a_i \leq c_1 < \dots < c_k \leq \max a_i$;

Step 2: Priduživanje (assignment step)

$$\pi_j = \{a_i \in \mathcal{A} : d(c_j, a_i) \leq d(c_s, a_i), \quad s = 1, \dots, k\}, \quad j = 1, \dots, k;$$

Step 3: Korekcija (update step)

$$\zeta_j = \operatorname{argmin}_{x \in \mathbb{R}} \sum_{a_i \in \pi_j} w_i d(x, a_i), \quad j = 1, \dots, k;$$

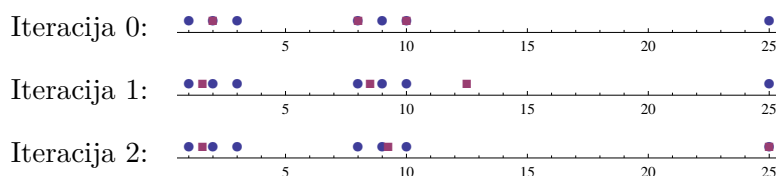
Step 2 i Step 3 se izmjenjuju tako dugo dok se ili centri ne poklope ili dok se particije ne poklope ili dok vrijednost funkcije cilja ne prestane opadati.

Primjer 10. Podacima iz Primjera 9 dodajmo težine kako slijedi: $w_i : 1, 1, .1, .5, .5, 1, .2$. Grupirat ćemo skup A u $k = 3$ klastera primjenom k-means algoritma uz početne centre kao i ranije. Rezultati su vidljivi u niže navedenoj tablici.

R.br.	z_1	z_2	z_3	F	π_1	π_2	π_3	ζ_1	ζ_2	ζ_3	\mathcal{F}
1	2	8	10	46.6	{1, 2, 3}	{8, 9}	{10, 25}	1.57143	8.5	12.5	38.4643
2	1.57143	8.5	12.5	34.4643	{1, 2, 3}	{8, 9, 10}	{25}	1.57143	9.5	25	2.08929
3	1.57143	9.5	25	2.08929	{1, 2, 3}	{8, 9, 10}	{25}	1.57143	9.5	25	2.08929

Prosječno težinsko kvadratno rasipanje po klasterima (varijanca): {0.340136, 0.6875, 0.}

Prosječno težinsko rasipanje po klasterima (standardna devijacija): {0.583212, 0.829156, 0.}



Slika 2: k-means algoritam za traženje lokalno optimalne particije skupa \mathcal{A} iz Primjera 10

4 Grupiranje na osnovi dva obilježja

Neka je $\mathcal{A} = \{a_1, \dots, a_m\}$ skup, koji treba na osnovi dva obilježja grupirati u k klastera koji zadovoljavaju uvjete iz Definicije 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 i jednakih.

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

$$\mathbf{c}_j = c(\pi_j) := \operatorname{argmin}_{\mathbf{x} \in \mathbb{R}^2} \sum_{\mathbf{a}_i \in \pi_j} d(\mathbf{x}, \mathbf{a}_i). \quad (29)$$

Kod problema grupiranja podataka u općem slučaju najčešće korištene kvazimetričke funkcije $d: \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}_+$ su [Ben-Israel and Iyigun, 2008, Gan et al., 2007, Kogan, 2007]

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2 \quad (\textit{least squares udaljenost}) \quad (30)$$

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_1 \quad (\textit{Manhattan udaljenost}) \quad (31)$$

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

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

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 $\mathcal{F}: \mathcal{P}(\mathcal{A}, k) \rightarrow \mathbb{R}_+$,

$$\mathcal{F}(\Pi) = \sum_{j=1}^k \sum_{\mathbf{a}_i \in \pi_j} d(\mathbf{c}_j, \mathbf{a}_i), \quad (32)$$

a d -optimalnu particiju Π^* tražimo rješavanjem optimizacijskog problema

$$\mathcal{F}(\Pi^*) = \min_{\Pi \in \mathcal{P}(\mathcal{A}, k)} \mathcal{F}(\Pi). \quad (33)$$

Pri tome, funkcija cilja (32) također može imati više lokalnih minimuma, koje također možemo tražiti primjenom Algoritma 2.

Primijetite da na taj način optimalna particija Π^* 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.

Obratno, za dani skup centara $c_1, \dots, c_k \in \mathbb{R}^2$, uz primjenu *principa minimalnih udaljenosti* možemo definirati particiju $\Pi = \{\pi_1, \dots, \pi_k\}$ skupa \mathcal{A} na sljedeći način:

$$\pi_j = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a), \forall s = 1, \dots, k\}, \quad j = 1, \dots, k, \quad (34)$$

pri čemu treba voditi računa o tome da svaki element skupa \mathcal{A} pripadne samo jednom klasteru.

Voronoijev dijagram

Zato se problem traženja optimalne particije skupa \mathcal{A} može svesti na sljedeći optimizacijski problem

$$\min_{c_1, \dots, c_k \in \mathbb{R}} F(c_1, \dots, c_k), \quad F(c_1, \dots, c_k) = \sum_{i=1}^m \min_{j=1, \dots, k} d(c_j, a_i), \quad (35)$$

gdje je $F: \mathbb{R}^{2k} \rightarrow \mathbb{R}_+$. Općenito, ova funkcija nije konveksna ni diferencijabilna, a može imati više lokalnih minimuma [Gan et al., 2007, Iyigun and Ben-Israel, 2010, Teboulle, 2007].

Optimizacijski problem (35) u literaturi se može naći pod nazivom *k-median problem* i ekvivalentan je optimizacijskom problemu (33). Naime, vrijedi

$$\begin{aligned} F(c_1, \dots, c_k) &:= \sum_{i=1}^m \min\{d(c_1, a_i), \dots, d(c_k, a_i)\} \\ &\geq \sum_{j=1}^k \sum_{a_i \in \pi_j} \min\{d(c_1, a_i), \dots, d(c_k, a_i)\} \\ &= \sum_{j=1}^k \sum_{a_i \in \pi_j} d(c_j, a_i) =: \mathcal{F}(\Pi), \end{aligned}$$

gdje je $\Pi = \{\pi_1, \dots, \pi_k\}$,

$$\pi_j = \pi_j(c_1, \dots, c_k) = \{a \in \mathcal{A} : d(c_j, a) \leq d(c_s, a), \forall s = 1, \dots, k\},$$

odnosno

$$\pi_j = \pi_j(c_1, \dots, c_k) = \{a_i \in \mathcal{A} : j = \operatorname{argmin}_{s=1, \dots, k} d(c_s, a_i)\},$$

pri čemu jednakost vrijedi ako je Π lokalno ili globalno optimalna particija.

4.1 Princip najmanjih kvadrata

Definicija 2. Neka je $\mathcal{A} = \{\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \dots, m\}$ skup vektora iz \mathbb{R}^2 . Kažemo da je particija $\Pi^* = \{\pi_1^*, \dots, \pi_k^*\}$ optimalna u smislu najmanjih kvadrata (skraćeno: LS-optimalna) ako je kvazimetrička funkcija $d: \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}_+$ definirana s (30), a Π^* rješenje optimizacijskog problema (32)–(33).

Primijetite da funkcija (30) nije metrika jer ne zadovoljava nejednakost trokuta. Centri $\mathbf{c}_1, \dots, \mathbf{c}_k$ klastera π_1, \dots, π_k određeni su s

$$\mathbf{c}_j = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^2} \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{a}_i - \mathbf{u}\|_2^2 = \frac{1}{|\pi_j|} \sum_{\mathbf{a}_i \in \pi_j} \mathbf{a}_i, \quad j = 1, \dots, k, \quad (36)$$

a funkcija cilja (32) s

$$\mathcal{F}(\Pi) = \sum_{j=1}^k \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{c}_j - \mathbf{a}_i\|_2^2 \quad (37)$$

Primjer 11. 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 Definiciju 1, a nakon toga odgovarajuće centroide i vrijednosti funkcije cilja (37) u smislu LS-optimalnosti.

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	$\begin{pmatrix} 2 \\ 3, 3 \end{pmatrix}$	$0 + \frac{4}{3} \approx 1.3$	$\frac{1}{2} + \frac{1}{6} \approx 0.6$
$\{\mathbf{a}_2\}$	$\{\mathbf{a}_1, \mathbf{a}_3, \mathbf{a}_4\}$	\mathbf{a}_2	$\begin{pmatrix} 1 \\ 3, 3 \end{pmatrix}$	$0 + \frac{4}{3} \approx 1.3$	$\frac{1}{2} + \frac{1}{6} \approx 0.6$
$\{\mathbf{a}_3\}$	$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_4\}$	\mathbf{a}_3	$\begin{pmatrix} 1 \\ 3, 3 \end{pmatrix}$	$0 + \frac{4}{3} \approx 1.3$	$\frac{1}{2} + \frac{1}{6} \approx 0.6$
$\{\mathbf{a}_4\}$	$\{\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3\}$	\mathbf{a}_4	$\begin{pmatrix} 2 \\ 3, 3 \end{pmatrix}$	$0 + \frac{4}{3} \approx 1.3$	$\frac{1}{2} + \frac{1}{6} \approx 0.6$
$\{\mathbf{a}_1, \mathbf{a}_2\}$	$\{\mathbf{a}_3, \mathbf{a}_4\}$	$\begin{pmatrix} 1 \\ 2, 0 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 2, 1 \end{pmatrix}$	$\frac{1}{2} + \frac{1}{2} = 1$	$\frac{1}{2} + \frac{1}{2} = 1$
$\{\mathbf{a}_1, \mathbf{a}_4\}$	$\{\mathbf{a}_2, \mathbf{a}_3\}$	$\begin{pmatrix} 0 \\ 1, 2 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 1, 2 \end{pmatrix}$	$\frac{1}{2} + \frac{1}{2} = 1$	$\frac{1}{2} + \frac{1}{2} = 1$
$\{\mathbf{a}_1, \mathbf{a}_3\}$	$\{\mathbf{a}_2, \mathbf{a}_4\}$	$\begin{pmatrix} 1 \\ 2, 2 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 2, 2 \end{pmatrix}$	$1 + 1 = 2$	$0 + 0 = 0$

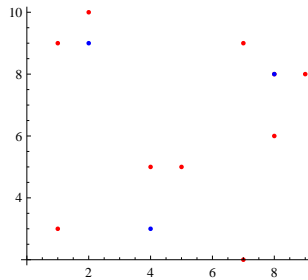
Broj svih dvočlanih particija ovog skupa je $2^{m-1} - 1 = 7$, a kao što se vidi iz tablice, dvije particije $\{\{\mathbf{a}_1, \mathbf{a}_2\}, \{\mathbf{a}_3, \mathbf{a}_4\}\}$ i $\{\{\mathbf{a}_1, \mathbf{a}_4\}, \{\mathbf{a}_2, \mathbf{a}_3\}\}$ su optimalne jer na njima kriterijska funkcija cilja (37) postiže globalni minimum (vidi sliku).

Primjer 12. Zadan je skup $\mathcal{A} = \{\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \dots, m\}$

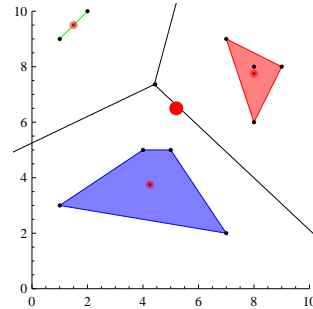
i	1	2	3	4	5	6	7	8	9	10
x_i	5	1	2	7	8	1	7	4	8	9
y_i	5	3	10	2	8	9	9	5	6	8

Uz početne centre $c_1 = (2, 9)$, $c_2 = (4, 3)$, $c_3 = (8, 8)$, primjenom LS-kvazimetričke funkcije i Algoritma 2 dobivamo lokalno optimalnu particiju $\Pi = \{\pi_1, \pi_2, \pi_3\}$ s centrima $c_1^* = (\frac{3}{2}, \frac{19}{2})$, $c_2^* = (\frac{17}{4}, \frac{15}{4})$, $c_3^* = (8, \frac{31}{4})$ (vidi Sliku 3). Vrijednost funkcije cilja je $\mathcal{F} = \frac{133}{4} = 1 + \frac{51}{2} + \frac{27}{4}$.

(a) Početna particija



(a) Optimalna particija



Slika 3: k -means algoritam za traženje lokalno optimalne particije skupa \mathcal{A} iz Primjera 12

4.1.1 Dualni problem

Analogno, kao u jednodimenzionalnom slučaju može se pokazati da vrijedi

$$\sum_{i=1}^m \|\mathbf{a}_i - \mathbf{c}\|_2^2 = \sum_{j=1}^k \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{c}_j - \mathbf{a}_i\|_2^2 + \sum_{j=1}^k m_j \|\mathbf{c}_j - \mathbf{c}\|_2^2, \quad (38)$$

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 (37) optimalnu LS-particiju možemo tražiti maksimizacijom funkcije

$$\mathcal{G}(\Pi) = \sum_{j=1}^k m_j \|\mathbf{c}_j - \mathbf{c}\|_2^2. \quad (39)$$

Određenim prilagodavanjem [Dhillon et al., 2004] problem se svodi na poznate probleme i metode linearne algebre.

Primjer 13. Skup \mathcal{A} iz Primjera 11 ima 7 različitih particija i za sve njih u tablici je prikazana vrijednost kriterijske funkcije cilja \mathcal{G} . Kao što se vidi, funkcija \mathcal{G} prima maksimalnu vrijednost na 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 Princip najmanjih apsolutnih odstupanja

Definicija 3. Neka je $\mathcal{A} = \{\mathbf{a}_i = (x_i, y_i) \in \mathbb{R}^2 : i = 1, \dots, m\}$ skup vektora iz \mathbb{R}^2 . Kažemo da je particija $\Pi^* = \{\pi_1^*, \dots, \pi_k^*\}$ optimalna u smislu najmanjih apsolutnih odstupanja (skraćeno: LAD-optimalna) ako je kvazimetrička funkcija $d : \mathbb{R}^2 \times \mathbb{R}^2 \rightarrow \mathbb{R}_+$ definirana s (31), a Π^* rješenje optimizacijskog problema (32)–(33).

Centri $\mathbf{c}_1, \dots, \mathbf{c}_k$ klastera π_1, \dots, π_k određeni su s

$$\mathbf{c}_j = \operatorname{argmin}_{\mathbf{u} \in \mathbb{R}^2} \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{a}_i - \mathbf{u}\|_1 = (\operatorname{med}(x), \operatorname{med}(y)) =: \operatorname{med}(\pi_j), \quad j = 1, \dots, k, \quad (40)$$

gdje je $x = (x_1, \dots, x_n)$, $y = (y_1, \dots, y_m) \in \mathbb{R}^m$, a funkcija cilja (32) zadana je s

$$\mathcal{F}(\Pi) = \sum_{j=1}^k \sum_{\mathbf{a}_i \in \pi_j} \|\mathbf{c}_j - \mathbf{a}_i\|_1 \quad (41)$$

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 [Vazler et al., 2012].

4.3 Primjena Mahalanobis udaljenosti

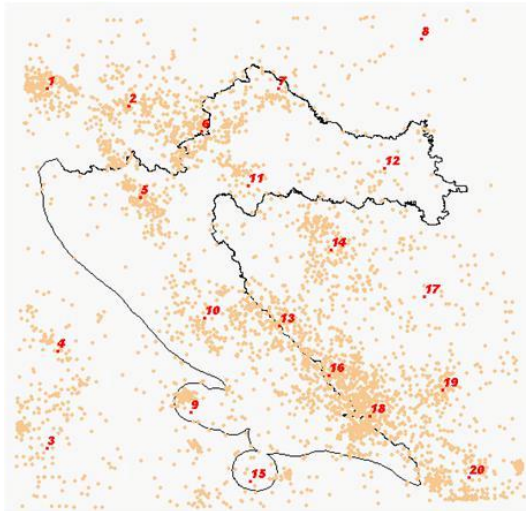
Vidi nastavni materijal [Reprezentant](#).

5 Neke primjene

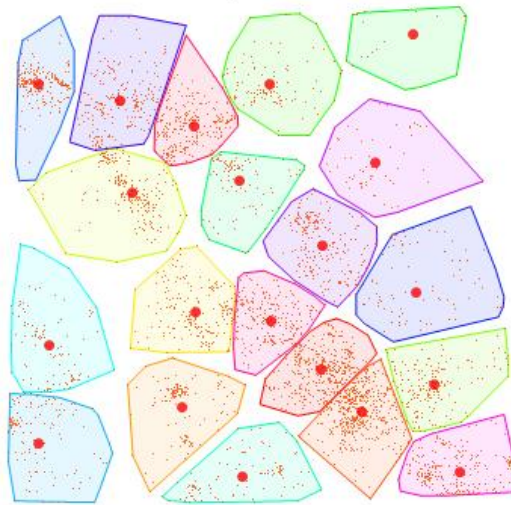
5.1 Detekcija centara seizmičkih aktivnosti

Reference: [Cho et al., 2010, Colombo et al., 1997, Stipčević et al., 2011]

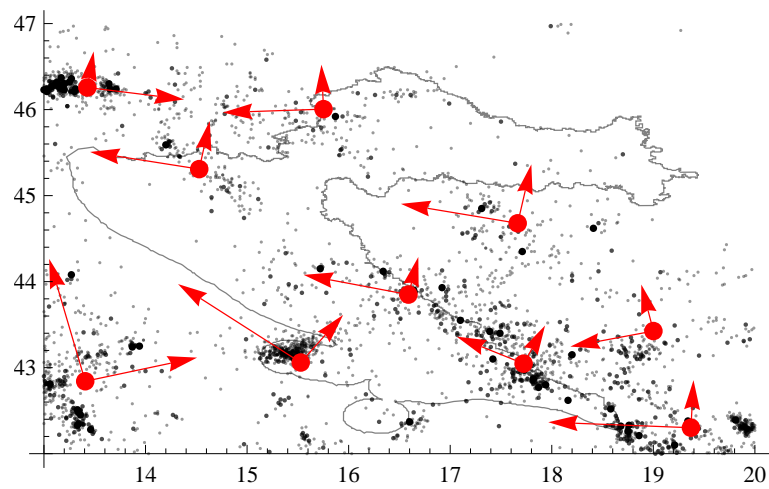
a) Lokacije potresa

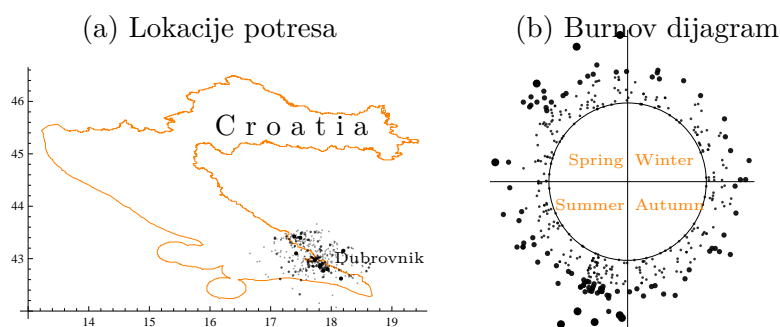


b) k=20



WLS-centers	WM-centers	N_5/N_{WLS}	N_5/N_{WM}	Location	Position
(13.39, 42.88)	(13.41, 42.84)	10/419	10/413	Ascoli Piceno (Italy)	(13.57, 42.85)
(13.35, 46.23)	(13.43, 46.26)	20/336	20/345	Udine (Italy)	(13.23, 46.08)
(14.74, 45.64)	(14.53, 45.31)	3/269	3/176	Kočevje (Slovenia)	(14.86, 45.63)
(15.47, 43.14)	(15.53, 43.06)	3/532	4/546	Šibenik (Croatia)	(15.88, 43.72)
(16.21, 45.89)	(15.75, 46.01)	3/129	3/210	Zagreb (Croatia)	(16.02, 45.82)
(16.63, 43.83)	(16.59, 43.85)	6/232	6/287	Sinj (Croatia)	(16.64, 43.70)
(17.59, 43.03)	(17.72, 43.05)	13/437	11/402	Metković (Croatia)	(17.64, 43.05)
(17.69, 44.64)	(17.66, 44.68)	4/186	4/180	Banja Luka (Bosnia and Herzegovina)	(17.19, 44.75)
(18.94, 43.24)	(19.00, 43.43)	2/220	1/163	Žabljak (Montenegro)	(19.11, 43.15)
(19.42, 42.26)	(19.37, 42.30)	9/424	11/462	Podgorica (Montenegro)	(19.27, 42.41)





Slika 4: Seizmička aktivnost u području Dubrovnika od 1973.

	$\{c_1\}$	$\{c_1, c_2\}$	$\{c_1, c_2, c_3\}$	$\{c_1, c_2, c_3, c_4\}$
Alg. 2	$\{4.15\}$	$\{1.67, 4.15\}$	$\{1.90, 4.46, 6.09\}$	$\{2.29, 2.52, 4.38, 6.24, \}$
Obj.Fun.	1922	1032	726	554
Alg. 1	-	$\{1.90, 4.46\}$	$\{2.29, 4.38, 6.24\}$	$\{1.59, 2.60, 4.38, 6.09\}$
Obj.Fun.	-	964	638	517
dd:mm	/ 05:04 /	/ 19:03 / 16:08 /	/ 11:04 / 11:08 / 28:11 /	/ 01:03 / 29:04 / 11:08 / 19:11 /

Tablica 2: Trenuci najintenzivnije seizmičke aktivnosti u godini

5.2 Istraživanje i predviđanje mogućnosti poplave na nekom mjestu riječnog toka

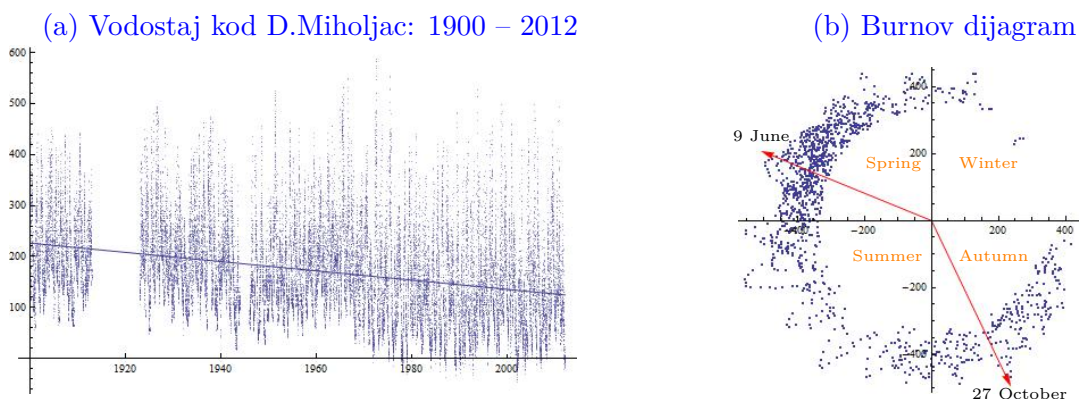
Reference: [Parajka et al., 2010]

$(T_i, w_i), i = 1, \dots, N, T_i \in [0, 112]$ (podaci između 1900-1-1 i 2012-2-1)

w_i (izmjereni vodostaj na dan T_i)

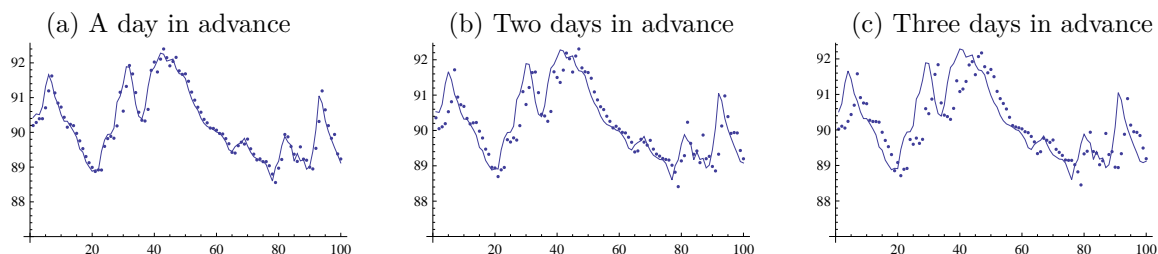
$t_i = 2\pi T_i \pmod{2\pi} \in [0, 2\pi]$, (transformirani datumi)

$$\mathcal{A} = \{a_i = w_i(\cos t_i, \sin t_i) \in \mathbb{R}^2 : w_i - l(t_i) \geq 100\}$$

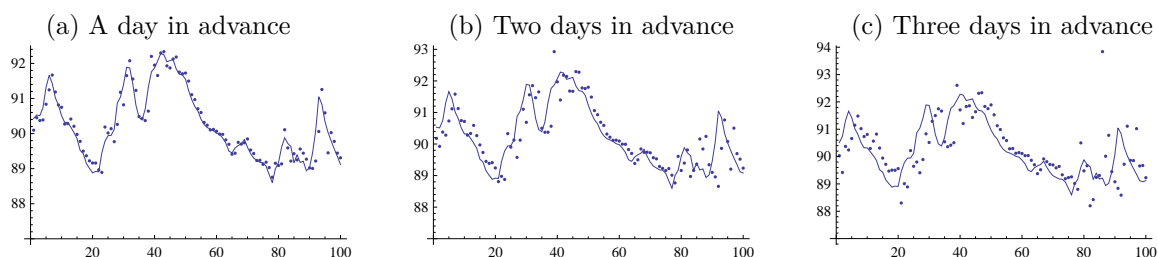


5.3 Kratkoročna prognoza vremenskih nizova

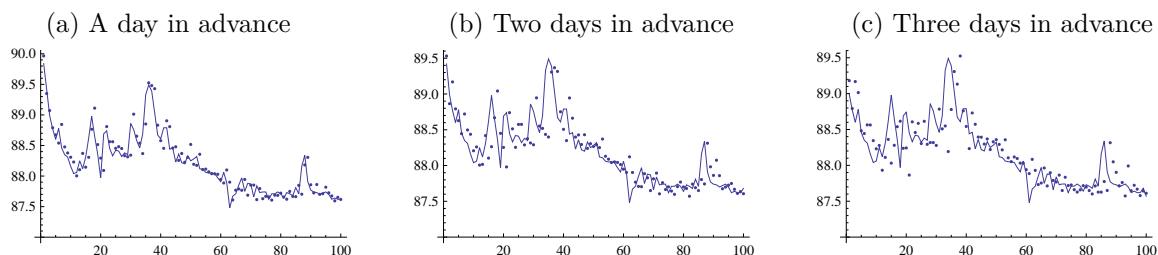
Reference: [Parajka et al., 2010, Wu et al., 2008]



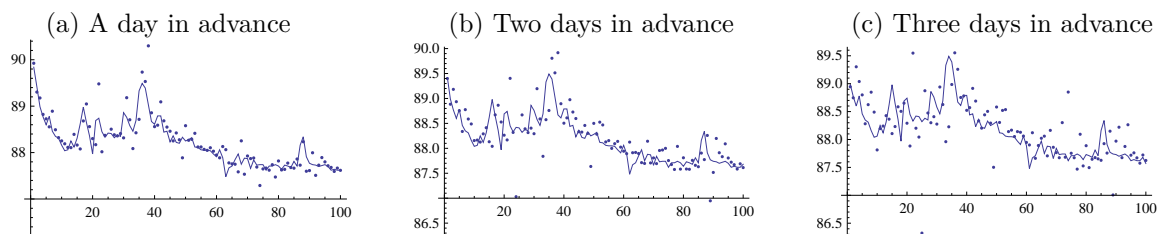
Slika 5: LAD water level prediction from 5 October 2000 to 12 January 2001



Slika 6: LS water level prediction from 5 October 2000 to 12 January 2001



Slika 7: LAD water level prediction from 18 September 2011 to 26 December 2011



Slika 8: LS water level prediction from 18 September 2011 to 26 December 2011

5.4 Predviđanje satne potrošnje prirodnog plina

Reference: [Sabo et al., 2011b]

5.5 Prihvatljivo definiranje izbornih jedinica

Reference: [Sabo et al., 2012]

5.6 Upravljanje pokretnim objektima bez prisutnosti čovjeka

Reference: V. Leemans, M.-F. Destain, *Line cluster detection using a variant of the Hough transform for culture row localisation*, Image and Vision Computing 24(2006) 541–55
 G.Hamarneh, K.Althoff, R. Abu-Gharbieh, *Automatic Line Detection with Hough Transform*, Project Report for the Computer Vision Course, Lund, May 5-7 and August 19-20, 1999

5.7 Prepoznavanje govora

M.El-Wakdy, E.El-Sehely, M.El-Tokhy, A.El-Hennawy, *Speech recognition using a wavelet transform to establish fuzzy inference system through subtractive clustering and neural network (ANFIS)* 12th WSEAS Int. Conf. on SYSTEMS, Heraklion, Greece, 22-24 June 2008

5.8 Pretraživanje teksta

Primjer 14. U nekom tekst prisutnost neke riječi kodira se s 1, a odsutnost te riječi iz teksta s 0. Postavlja se pitanje o sličnosti/različitosti dva teksta obzirom na prisutnost/odsutnost promatranih riječi. Tekst u kome je prisutno/odsutno $n \geq 1$ izabranih riječi prikazat ćemo vektorom iz \mathbb{R}^n s komponentama 0 ili 1. Primjerice, sljedećim vektorima prikazani su tekstovi u kojima su prisutne/odsutne riječi A,B,C:

$a^1 = (1, 1, 0)$: tekst u kome se pojavljuju riječi A,B, a ne pojavljuje riječ C

$a^2 = (1, 0, 0)$: tekst u kome se pojavljuje riječ A, a ne pojavljuje riječi B,C

$a^3 = (1, 0, 1)$: tekst u kome se pojavljuju riječi A, C, a ne pojavljuje riječ B

$a^4 = (0, 0, 1)$: rečenica u kojoj se pojavljuje riječ C, a ne pojavljuje riječi A, B

U svrhu ispitivanja sličnosti/različitosti tekstova obzirom na prisutnost/odsutnost nekih riječi možemo pokušati iskoristiti ranije spomenute metričke funkcije d_1, d_2, d_∞ . U znanstvenoj literaturi (vidi primjerice [Berry et al., 1999, Chaovalitwongse et al., 2009]) u tu svrhu koriste se neke tzv. kvazimetričke funkcije, kao što su

$d_{LS}(x, y) = \|x - y\|^2$ – Least Squares (LS) kvazimetrička funkcija

$d_c(x, y) = 1 - \frac{\langle x, y \rangle}{\|x\| \cdot \|y\|}$ – kosinus kvazimetrička funkcija

Od kvazimeričkih funkcija zahtijeva se samo svojstvo pozitivne definitnosti, dok ostala svojstva metričkih funkcija ne moraju biti ispunjena. Ipak prethodno spomenute dvije kvazimeričke funkcije, pored svojstva pozitivne definitnosti, zadovoljavaju i svojstvo simetričnosti, ali ne zadovoljavaju nejednakost trokuta. Za prethodno spomenuti primjer dobivamo

$$\begin{array}{lll} d_{LS}(a^1, a^2) = 1, & d_{LS}(a^1, a^3) = 2, & d_{LS}(a^1, a^4) = 3 \\ d_1(a^1, a^2) = 1, & d_1(a^1, a^3) = 2, & d_1(a^1, a^4) = 3, \\ d_c(a^1, a^2) = 1 - \frac{\sqrt{2}}{2} = 0.29, & d_c(a^1, a^3) = \frac{1}{2}, & d_c(a^1, a^4) = 1 \end{array}$$

Prema LS-kvazimetričkoj funkciji (a također i prema l_1 -metričkoj funkciji) tekstovi a^1 i a^2 su najbliži (najbliži), a tekstovi a^1 i a^4 najrazličitiji (najudaljeniji) obzirom na pojavu riječi A,B,C.

I prema kosinus-metričkoj funkciji d_c tekstovi a^1 i a^2 su najbliži (najbliži), a tekstovi a^1 i a^4 potpuno različiti (maksimalno udaljeni) obzirom na pojavu riječi A,B,C.

Primjer 15. *Promatramo tekstove u kojima se mogu pojaviti riječi: A,B,C,D,E. Neka je primjerice:*

$a^1 = (1, 0, 0, 0, 1)$: tekst u kome se pojavljuju riječi A, E, a ne pojavljuju riječi B,C,D

$a^2 = (0, 1, 1, 0, 0)$: tekst u kome se pojavljuju riječi B, C, a ne pojavljuju riječi A,D,E

$a^3 = (1, 0, 0, 0, 0)$: tekst u kome se pojavljuje riječ A, a ne pojavljuju riječi B,C,D, E

$d_{LS}(a^i, a^j)$	a^1	a^2	a^3	$d_1(a^i, a^j)$	a^1	a^2	a^3	$d_c(a^i, a^j)$	a^1	a^2	a^3
a^1	0	4	1	a^1	0	4	1	a^1	0	1	0.29
a^2	4	0	3	a^2	4	0	3	a^2	1	0	1
a^3	1	3	0	a^3	1	3	0	a^3	0.29	1	0

Iz ovog primjera vidi se da kosinus-kvazimetrička funkcija puno bolje identificira sličnosti/različitosti tekstova obzirom na prisutnost/odsutnost riječi A,B,C,D,E (objasnite to na osnovi brojeva iz tablica!).

Literatura

- F. AURENHAMMER, R. KLEIN, *Voronoi diagrams*, In: J. SACK, G. URRUTIA, editors, *Handbook of Computational Geometry, Chapter V*. Elsevier Science Publishing, 2000, 201–290.
- M. S. BAZARAA, H. D. SHERALI, C. M. SHETTY, *Nonlinear Programming: Theory and Algorithms, 3rd Edition*, Wiley, New Jersey, 2006.
- A. BEN-ISRAEL, C. IYIGUN, *Probabilistic d-clustering*, Journal of Classification, **25**(2008) 5–26, URL <http://dx.doi.org/10.1007/s00357-008-9002-z>, 10.1007/s00357-008-9002-z.
- M. W. BERRY, Z. DRMAČ, E. R. JESSUP, *Using linear algebra for information retrieval*, SIAM Review, **41**(1999) 335–362.
- D. L. BOLEY, *Principal direction divisive partitioning*, Data Mining and Knowledge Discovery, **2**(1998) 325–344.
- J. F. BONNANS, J. C. GILBERT, C. LEMARECHAL, C. A. SAGASTIZABAL, *Numerical Optimization. Theoretical and Practical Aspects*, Springer-Verlag, Berlin, 2006.
- D. L. BOYD, L. VANDENBERGHE, *Convex optimization*, Cambridge University Press, Cambridge, 2004.
- J. BRIMBERG, H. JUEL, A. SCHÖBEL, *Properties of three-dimensional median line location models*, Annals of Operations Research, **122**(2003) 71–85.
- J. A. CADZOW, *Minimum l_1 , l_2 and l_∞ norm approximate solutions to an overdetermined system of linear equations*, Digital Signal Processing, **12**(2002) 524–560.

- W. A. CHAOVALITWONGSE, S. BUTENKO, P. M. PARDALOS, editors, *Clustering Challenges in Biological Networks*. World Scientific, 2009.
- N. F. CHO, K. F. TIAMPO, S. D. MCKINNON, J. A. VALLEJOS, W. KLEIN, , R. DOMINGUEZ, *A simple metric to quantify seismicity clustering*, *Nonlin. Processes Geophys.*, **17**(2010) 293–302.
- D. COLOMBO, V. GITIS, R. DE FRANCO, *Application of pattern recognition techniques to long-term earthquake prediction in central costa rica*, *Engineering Geology*, **48**(1997) 7–18.
- A. CORD, C. AMBROISE, J.-P. COCQUEREZ, *Feature selection in robust clustering based on laplace mixture*, *Pattern Recognition Letters*, **27**(2006) 627–635.
- R. CUPEC, R. GRBIĆ, K. SABO, R. SCITOVSKI, *Three points method for searching the best least absolute deviations plane*, *Applied Mathematics and Computation*, **215**(2009) 983–994.
- D. DAVIES, D. BOULDIN, *A cluster separation measure*, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **2**(1979) 224–227.
- W. H. E. DAY, H. EDELSBRUNNER, *Efficient algorithms for agglomerative hierarchical clustering methods*, *Journal of Classification*, **1**(1984) 7–24.
- E. DEMIDENKO, *Criteria for unconstrained global optimization*, *J. Optim. Theory Appl.*, **136**(2008) 375–395.
- J. J. DENNIS, R. SCHNABEL, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.
- I. S. DHILLON, Y. GUAN, B. KULIS, *Kernel k-means, spectral clustering and normalized cuts*, In: *Proceedings of the 10-th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), August 22–25, 2004, Seattle, Washington, USA*. 2004, 551–556.
- D. DI SERAFINO, G. LIUZZI, V. PICCIALI, F. RICCIO, G. TORALDO, *A modified dividing rectangles algorithm for a problem in astrophysics*, *J Optim Theory Appl*, **151**(2011) 175–190.
- Y. DODGE, *An introduction to l_1 -norm based statistical data analysis*, *Computational Statistics & Data Analysis*, **5**(1987) 239–253.
- B. DURAK, *A Classification Algorithm Using Mahalanobis Distances Clustering of Data with Applications on Biomedical Data Set*, Ph.D. thesis, The Graduate School of Natural and Applied Sciences of Middle East Technical University, 2011.
- B. S. EVERITT, S. LANDAU, M. LEESE, *Cluster analysis*, Wiley, London, 2001.
- Y. G. EVTUSHENKO, *Numerical Optimization Techniques (Translations Series in Mathematics and Engineering)*, Springer-Verlag, Berlin, 1985.
- D. E. FINKEL, *DIRECT Optimization Algorithm User Guide*, Center for Research in Scientific Computation. North Carolina State University, 2003, <http://www4.ncsu.edu/definkel/research/index.html>.
- C. A. FLOUDAS, C. E. GOUNARIS, *A review of recent advances in global optimization*, *Journal of Global Optimization*, **45**(2009) 3–38.

- J. M. GABLONSKY, *Direct version 2.0*, Technical report, Center for Research in Scientific Computation. North Carolina State University, 2001.
- G. GAN, C. MA, J. WU, *Data Clustering: Theory, Algorithms, and Applications*, SIAM, Philadelphia, 2007.
- C. GURWITZ, *Weighted median algorithms for l_1 approximation*, BIT, **30**(1990) 301–310.
- C. IYIGUN, *Probabilistic distance clustering*, Ph.D. thesis, Graduate School – New Brunswick, Rutgers, 2007.
- C. IYIGUN, A. BEN-ISRAEL, *Contour approximation of data: A duality theory*, Linear Algebra and its Applications, **430**(2009) 2771–2780.
- C. IYIGUN, A. BEN-ISRAEL, *A generalized weiszfeld method for the multi-facility location problem*, Operations Research Letters, **38**(2010) 207–214.
- D. R. JONES, C. D. PERTTUNEN, B. E. STUCKMAN, *Lipschitzian optimization without the Lipschitz constant*, Journal of Optimization Theory and Applications, **79**(1993) 157–181.
- C. S. KAPLAN, *Voronoi diagrams and ornamental design*, In: N. A. FRIEDMAN, J. BARRALO, editors, *Proceedings of ISAMA '99*. 1999, 277–283.
- L. KAUFMAN, P. J. ROUSSEEUW, *Finding groups in data: An introduction to cluster analysis*, John Wiley & Sons, Hoboken, 2005.
- J. KOGAN, *Introduction to Clustering Large and High-dimensional Data*, Cambridge University Press, 2007.
- J. KOGAN, M. TEBoulLE, *Scaling clustering algorithms with bregman distances*, In: M. W. BERRY, M. CASTELLANOS, editors, *Proceedings of the Workshop on Text Mining at the Sixth SIAM International Conference on Data Mining*. 2006.
- D. E. KVASOV, Y. D. SERGEYEV, *A univariate global search working with a set of lipschitz constants for the first derivative*, Optimization Letters, **3**(2009) 303–318.
- D. E. KVASOV, Y. D. SERGEYEV, *Univariate geometric lipschitz global optimisation algorithms*, Numerical Algebra, Control and Optimization, **2**(2012) 69–90.
- F. LEISCH, *A toolbox for k -centroids cluster analysis*, Computational Statistics & Data Analysis, **51**(2006) 526–544.
- J. MOCKUS, *On the pareto optimality in the context of lipschitzian optimization*, Informatica, **22**(2011) 524–536.
- A. NEUMAIER, *Complete search in continuous global optimization and constraint satisfaction*, Acta Numerica, **13**(2004) 271 – 369.
- M. NG, *A note on constrained k -means algorithms*, Pattern Recognition, **33**(2000) 525–519.
- Y. NIEVERGELT, *Total least squares: state-of-the-art regression in numerical analysis*, SIAM Review, **36**(1994) 258–264.

- K. NYARKO, R. SCITOVSKI, *Solving the parameter identification problem of mathematical model using genetic algorithms*, Applied Mathematics and Computation, **153**(2004) 651–658.
- A. OKABE, B. BOOTS, K. SUGIHARA, *Spatial Tessellations: Concepts and Applications of Voronoi diagrams*, John Wiley & Sons, Chichester, UK, 2000.
- J. PARAJKA, S. KOHNOVÁ, G. BÁLINT, M. BARBUC, M. BORGA, P. CLAPS, S. C. A. DUMITRESCU, E. GAUME, K. HLAVČOVÁ, R. MERZ, M. PFAUNDLER, G. STANCALIE, J. SZOLGAY, G. BLÖSCHL, *Seasonal characteristics of flood regimes across the alpine–carpathian range*, Journal of Hydrology, **394**(2010) 78–89.
- P. M. PARDALOS, P. HANSEN, *Data mining and mathematical programming*, In: -. American Mathematical Society, Providence, 2008.
- S. A. PIJAVSKIJ, *An algorithm for searching for a global minimum of a function*, USSR Computational Mathematics and Mathematical Physics, **12**(1972) 888–896, (in Russian).
- P. J. ROUSSEEUW, A. M. LEROY, *Robust Regression and Outlier Detection*, Wiley, New York, 2003.
- A. RUSZCZYNSKI, *Nonlinear Optimization*, Princeton University Press, Princeton and Oxford, 2006.
- K. SABO, R. SCITOVSKI, *The best least absolute deviations line – properties and two efficient methods*, ANZIAM Journal, **50**(2008) 185–198.
- K. SABO, R. SCITOVSKI, I. VAZLER, *Data clustering (in croatian)*, Osječki matematički list, **10**(2010) 149–178.
- K. SABO, R. SCITOVSKI, P. TALER, *Uniform distribution of the number of voters per constituency on the basis of a mathematical model (in croatian)*, Hrvatska i komparativna javna uprava, **14**(2012) 229–249.
- K. SABO, R. SCITOVSKI, I. VAZLER, *Searching for a best LAD-solution of an overdetermined system of linear equations motivated by searching for a best LAD-hyperplane on the basis of given data*, Journal of Optimization Theory and Applications, **149**(2011a) 293–314.
- K. SABO, R. SCITOVSKI, I. VAZLER, *One-dimensional center-based l_1 -clustering method*, Optimization Letters, -(in press) DOI: 10.1007/s11590-011-0389-9.
- K. SABO, R. SCITOVSKI, I. VAZLER, M. ZEKIĆ-SUŠAC, *Mathematical models of natural gas consumption*, Energy Conversion and Management, **52**(2011b) 1721–1727.
- Y. D. SERGEYEV, D. E. KVASOV, *Lipschitz global optimization*, In: J. COCHRAN, editor, *Wiley Encyclopedia of Operations Research and Management Science*, volume 4. Wiley, New York, 2011, 2812–2828.
- B. SHUBERT, *A sequential method seeking the global maximum of a function*, SIAM Journal on Numerical Analysis, **9**(1972) 379–388.
- H. SPÄTH, *Cluster-Formation und Analyse*, R. Oldenburg Verlag, München, 1983.

- J. STIPČEVIĆ, H. TKALČIĆ, M. HERAK, S. MARKUŠIĆ, D. HERAK, *Crustal and uppermost mantle structure beneath the external dinarides, croatia, determined from teleseismic receiver functions*, *Geophys. J. Int.*, **185**(2011) 1103–1119.
- R. G. STRONGIN, Y. D. SERGEYEV, *Global Optimization with Non-Convex Constraints: Sequential and Parallel Algorithms*, Kluwer Academic Publishers, Dordrecht, 2000.
- Z. SU, J. KOGAN, C. NICHOLAS, *Constrained clustering with k-means type algorithms*, In: M. W. BERRY, J. KOGAN, editors, *Text Mining: Applications and Theory*. Wiley, Chichester, 2010, 81–103.
- M. TEBoulLE, *A unified continuous optimization framework for center-based clustering methods*, *Journal of Machine Learning Research*, **8**(2007) 65–102.
- R. J. VANDERBEI, *Extension of piyavskii's algorithm to continuous global optimization*, *Journal of Global Optimization*, **14**(1999) 205–216.
- F. P. VASILJEV, *Methods for Solving Extrema Problems*, University of Moscow, 1974, (in Russian).
- I. VAZLER, K. SABO, R. SCITOVSKI, *Weighted median of the data in solving least absolute deviations problems*, *Communications in Statistics - Theory and Methods*, **41:8**(2012) 1455–1465.
- L. VENDRAMIN, R. J. G. B. CAMPello, E. R. HRUSCHKA, *On the comparison of relative clustering validity criteria*, In: *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA. SIAM 2009*. SIAM, 2009, 733–744.
- V. VOLKOVICH, J. KOGAN, C. NICHOLAS, *Building initial partitions through sampling techniques*, *European Journal of Operational Research*, **183**(2007) 1097–1105.
- E. WEISZFELD, *Sur le point pour lequel la somme des distances de n points donnees est minimum*, *The Tohoku Mathematical Journal*, **43**(1936) 355–386.
- C. WU, K. CHAU, Y. LI, *River stage prediction based on a distributed support vector regression*, *Journal of Hydrology*, **358**(2008) 96– 111.
- X. S. YANG, *Firefly algorithms for multimodal optimization*, In: *Proceedings of the 5th international conference on Stochastic algorithms: foundations and applications*. 2009, 169–178.
- T. ZHANG, R. RAMAKRISHNAN, M. LIVNY, *BIRCH: an efficient data clustering method for very large databases*, In: *ACM SIGMOD Conference on Management of Data*. ACM, 1996, 103–114.
- S. ZLOBEC, *The fundamental theorem of calculus for lipschitz functions*, *Mathematical Communications*, **13**(2008) 215–232.