TAUTSAIMNIECĪBAS ATTĪSTĪBAS PROBLĒMAS UN RISINĀJUMI

Pēteris Grabusts

Rēzeknes Augstskola

2012. gada 19. aprīlis

KLASTERIZĀCIJU RAKSTUROJOŠO PARAMETRU IETEKME UZ DATU ANALĪZES REZULTĀTIEM

Uzdevumi

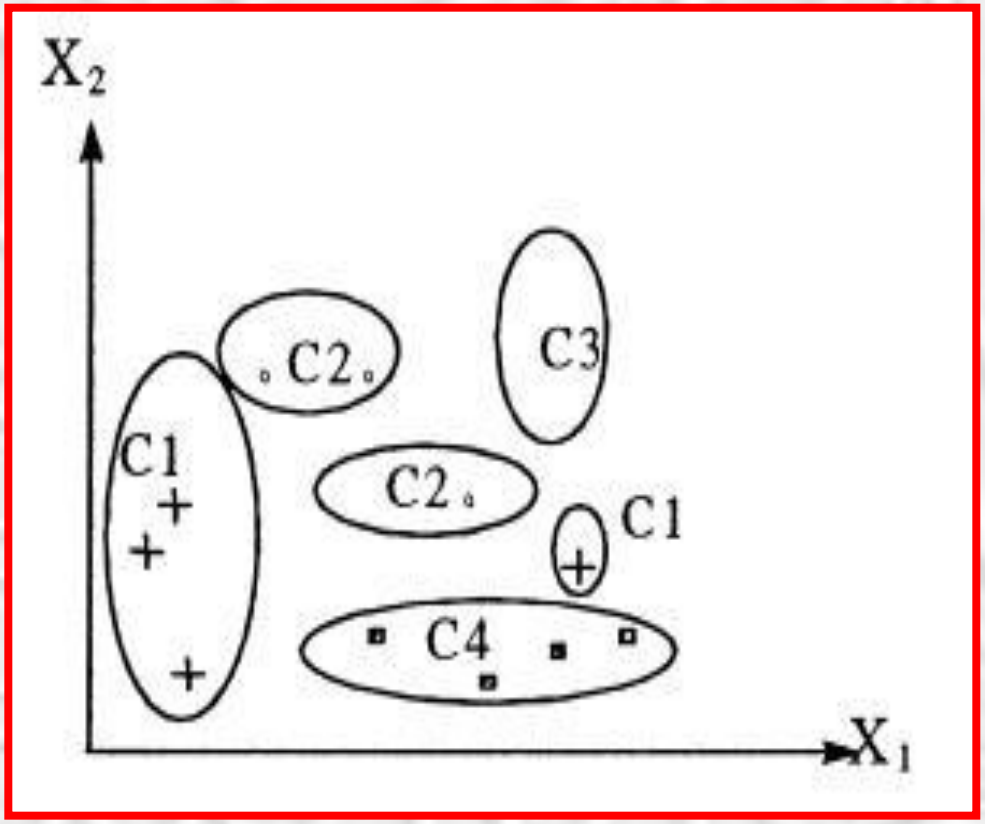
-**Metrikas izvēles pamatotība klasterizācijā**

-**Klasteru skaita izmaiņas pamatotība**

-**Klasterizācijas rezultātu ticamība**

-**Datu analīzes piemērs**

I. Klasiskās klasterizācijas metodes



Algoritmi -

- K-Means Clustering -

FOREL u.c.

- Mērķis -
  Ieejas vektorus sadalīt klasēs (klasteros) un noteikt to centrus

- Objektu grupas -
  klasteri, klases, taksoni

K-Means klasterizācijas algoritms

1 solis. Inicializē klasteru centrus w_j (j – nepieciešamo klasteru skaits uzdevuma risināšanai).

2.solis. Grupē visus apmācības izlases punktus ap tuvākā klastera centru t.i. katru punktu x_i saista ar klasteru j^*, kuram

$$\| x_i - w_{j^*} \| = \min_j \| x_i - w_j \|$$

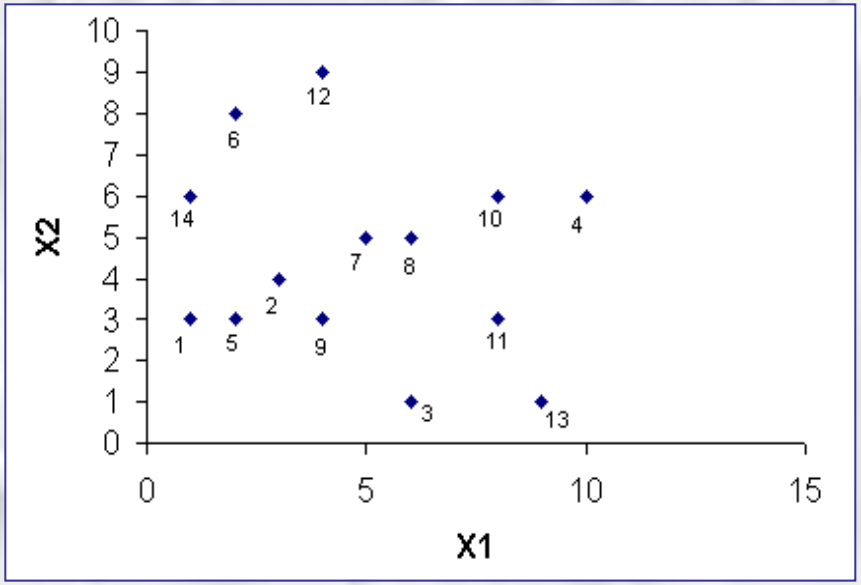
3.solis. Izskaitļo jaunos klasteru centrus, t.i. visiem w_j izskaitļo :

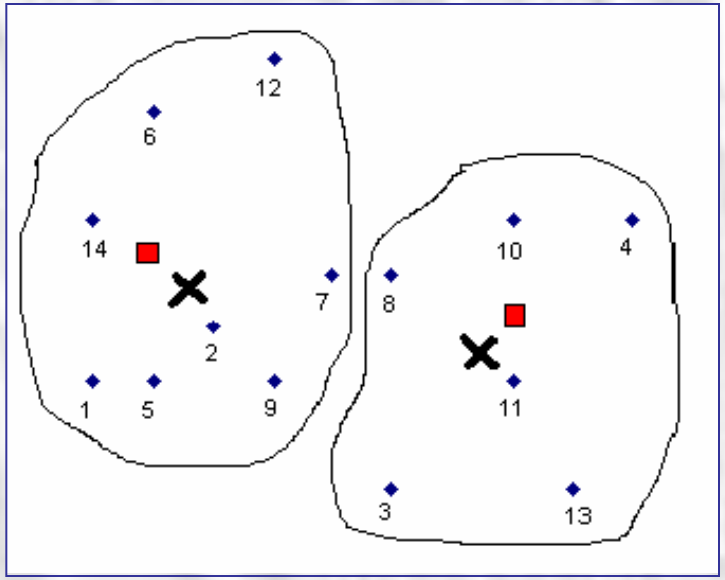
$$w_j = \frac{1}{m_j} \sum_{x_i \in klasteram\ j} x_i \quad ,$$

kur m_j – klasteram j piederošo punktu skaits.

4 solis. Atkārtot soli 2 tik ilgi, kamēr iterāciju laikā nemainās klasteru centru vērtības.

Klasterizācijas piemērs





Divi klasteri ar centriem punktos (-0.73; 0.26) un (0.97; -0.35)

rādiuss σ_1^2 = 1.07 un σ_2^2 = 1.04

Biežāk lietotā metrika - Eiklīda distance

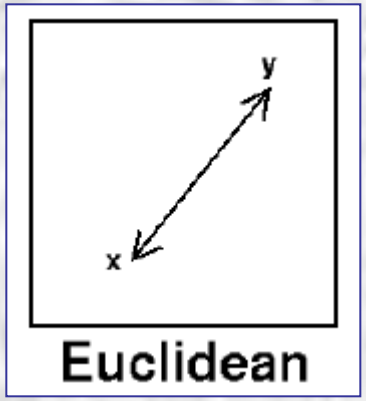
II. Metrika

Metrika jeb attālums ir funkcija, kas definē distanci starp datu elementiem

Metrika-1

Eiklīda distance ir visbiežāk izmantotā:

$$D_{XY} = \sqrt{\sum_{k=1}^{m} (x_{ik} - x_{jk})^2}$$



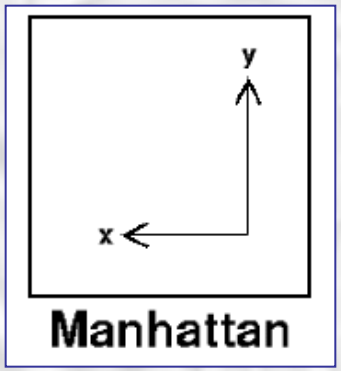
Euclidean distance between point X (1,2,3,4) and Y (5,6,7,8) is:

$$D_{XY} = \sqrt{(2-3)^2 + (3-5)^2 + (4-7)^2 + (5-9)^2} = 5{,}5$$

Metrika-2

Manhattan distance:

$$D_{XY} = \sum_{k=1}^{d} | x_{ik} - x_{jk} |$$

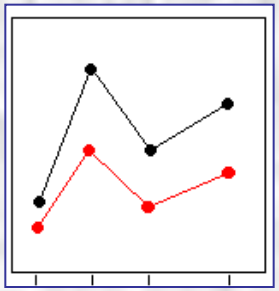


The Manhattan distance between point X (1,2,3,4) and Y (5,6,7,8) is:

$$D_{XY} = |2 - 3| + |3 - 5| + |4 - 7| + |5 - 9| = 10$$

Metrika-3

Korelācijas koeficients:



$$D_{XY} = (1 - r_{ij}) / 2$$

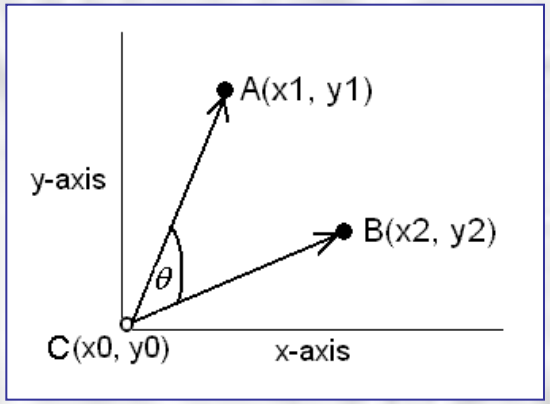
$$r_{ij} = \frac{\sum_{k=1}^{d} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{d} (x_{ik} - \bar{x}_i)^2 \sum_{k=1}^{d} (x_{jk} - \bar{x}_j)^2}}$$

The correlation coefficient between point X (1,2,3,4) and Y (5,6,7,8)

$$r_{XY} = \frac{(2-3{,}5)(3-6)+(3-3{,}5)(5-6)+(4-3{,}5)(7-6)(5-3{,}5)(9-6)}{([(2-3{,}5)^2+(3-3{,}5)^2+(4-3{,}5)^2+(5-3{,}5)^2][(3-6)^2+(5-6)^2+(7-6)^2+(9-6)^2])^{0{,}5}} = \frac{4{,}5+0{,}5+0{,}5+4{,}5}{\sqrt{5\times 20}} = 1$$

Metrika-4

Cosine distance:



$$D_{XY} = \cos(\theta) = \frac{X \bullet Y}{\|X\|\|Y\|} = \frac{\sum_{i=1}^{n} X_i \times Y_i}{\sqrt{\sum_{i=1}^{n} (X_i)^2} \times \sqrt{\sum_{i=1}^{n} (Y_i)^2}}$$

Pielietojumi

Measure	Examples and applications
Euclidean distance	K-means with its variants
Manhattan distance	Fuzzy ART, clustering algorithms
Chebyshev distance	Fuzzy C-means clustering
Minkowski distance	Fuzzy C-means clustering
Pearson correlation	Widely used as the measure for microarray gene expression data analysis

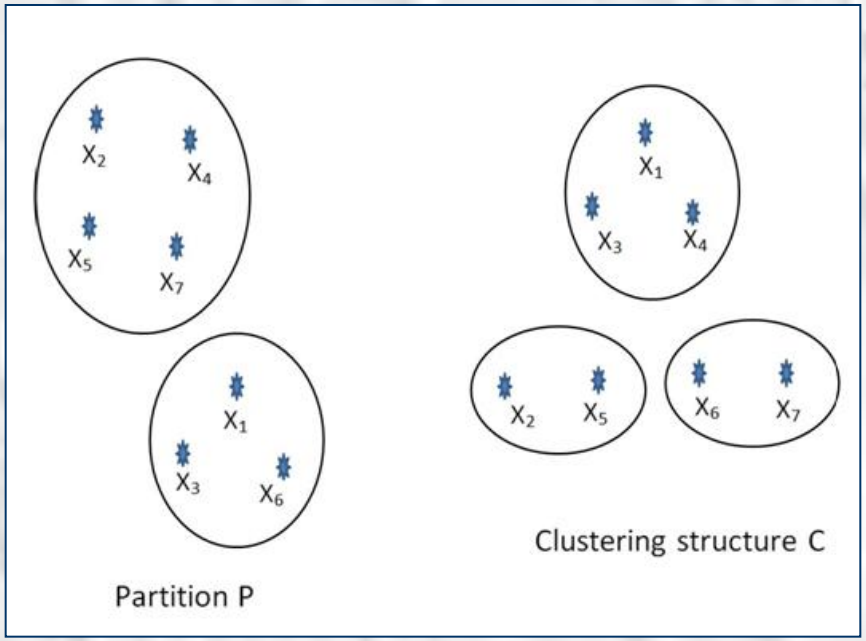
III. Klasterizācijas rezultātu ticamība

Skaitlisks kritērijs, kas ļauj novērtēt klasterizācijas rezultātu ticamību.

Trīs kritēriju grupas:

-ārējie kritēriji
-iekšējie kritēriji
-relatīvie kritēriji

Ārējā kritērija pieeja



Total number of pairs of points is

$$M = a + b + c + d = \frac{n(n-1)}{2}$$

where **n** is the number of data points in the data set.

Case	Pairs of data points	Total
a	x_1 and x_3; x_2 and x_5	2
b	x_1 and x_4; x_3 and x_4; x_6 and x_7	3
c	x_1 and x_6; x_2 and x_4; x_2 and x_7; x_3 and x_6; x_4 and x_5; x_4 and x_7; x_5 and x_7	7
d	x_1 and x_2; x_1 and x_5; x_1 and x_7; x_2 and x_3; x_2 and x_6; x_3 and x_5; x_3 and x_7; x_4 and x_6; x_5 and x_6	9

Kritēriju skaitliskā izteiksme

Rand index:

$$R = \frac{a+d}{M}$$

Jaccard coefficient:

$$J = \frac{a}{a+b+c}$$

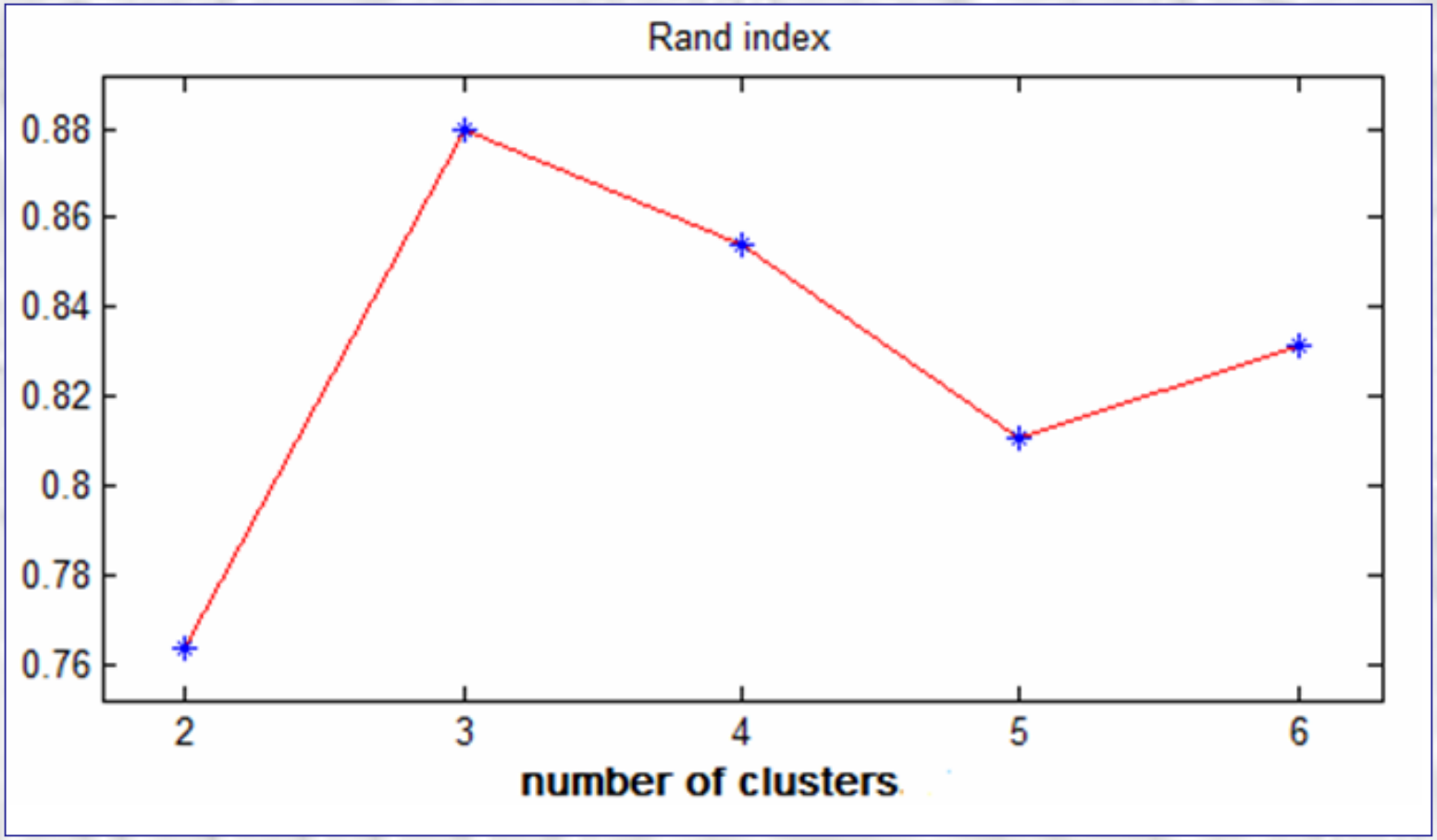
Fowlkes and Mallows index:

$$FM = \sqrt{\frac{a}{a+b}\frac{a}{a+c}}$$

Hubert's index:

$$\Gamma = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} X_{ij}Y_{ij}$$

Randa indekss



IV. Datu analīzes piemērs

Latvijas augstskolu 2011. gada reitinga analīze ar klasterizācijas palīdzību

Webometrics Ranking

Reitings *Webometrics Ranking* ranžē vairāk nekā 20 000 augstākās mācību iestādes pasaulē. Reitings balstās tikai uz Internetā pieejamās informācijas par mācību iestādi. Tiek izmantoti četri galvenie indikatori:

- 10% no ranga vērtības sastāda augstskolas atpazīstamība Google meklētājā;
- 50% - ārējo saišu skaits uz augstskolas mājas lapu;
- 10% - akadēmiskās un publicēšanās aktivitātes dažādu datņu formātā Google meklētājā (.doc, .pdf, .ppt);
- 30% - elektronisko publikāciju skaits no Google Scholar (2007-2011) un dati no Scimago SIR (2003-2010).

Šajā reitingā LU ierindota 822. vietā, RTU - 915. vietā, LLU - 3119. vietā, TSI - 3436. vietā, RA – 3659. vietā.

SCImago Institutions Rankings

Reitings *SCImago Institutions Rankings* ranžē 3042 augstākās mācību iestādes pasaulē un balstās uz datiem par augstskolas zinātniskajām aktivitātēm.
Četri indikatori ietver informāciju par
publikāciju skaitu (galvenokārt SCOPUS),
zinātniskās sadarbības rādītājus,
augsta līmeņa publikāciju skaitu.

No Latvijas augstskolām šeit minētas:
LU (1565. vietā)
RTU (2794. vietā).

QS World University Rankings

Reitings *QS World University Rankings* veido 500 pasaules vadošo augstskolu izlasi.
Tiek izmantoti 6 indikatori:
40% - akadēmiskā reputācija;
10% - darba devēju reputācija;
20% - zinātnisko darbu citējamība;
20% - studentu īpatsvars;
5% - ārzemju studentu īpatsvars;
5% - starptautisko fakultāšu īpatsvars.

Latvijas augstskolas šajā reitingu tabulā nav pārstāvētas.

The Times Higher World University Ranking (THE)

Reitings *The Times Higher World University Ranking (THE)* veido 400 pasaules vadošo augstskolu izlasi. Tiek izmantoti 13 indikatori, kas grupēti 5 grupās:

- 30% - apmācības vide;
- 30% - pētnieciskais darbs;
- 30% - citējamība;
- 2,5% - inovācijas;
- 7,5% - ārzemju sakari.

Latvijas augstskolas šajā reitingu tabulā nav pārstāvētas.

Latvijas augstskolu reitings

(jau 5 reizes)

Latvijas augstskolu reitinga izveidē pamatā ņemta THE metodoloģija un vērtēšanas kritēriji jeb indikatori ir šādi:

I1- studējošo un akadēmiskā personāla skaita attiecība;

I2- absolventu īpatsvars;

I3- akadēmiskā personāla pamatdarbā ar Dr. grādu īpatsvars (starp visām augstskolām);

I4- akadēmiskā personāla pamatdarbā ar Dr. grādu īpatsvars (konkrētajā augstskolā)

I5- akadēmiskā personāla pamatdarbā īpatsvars;

I6- akadēmiskā personāla vecuma struktūra (30 – 50 g. veco īpatsvars);

I7- ārzemju studentu īpatsvars;

I8- publikāciju skaits uz akadēmiskā personāla 1 vienību;

I9- izglītības kvalitāte (izcila un laba);

I10- augstskolas popularitāte/ atpazīstamība.

Latvijas Valsts augstskolu reitinga dati 2011.g.

Augst-skola	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Vieta
LU	63	13,5	150	63	44,5	46	39,5	200	184	100	1
RTU	45	11	124,5	60	50	36	21,5	166	200	100	2
RSU	53	12,5	57	57	50	38	44	100	198	99	3
DU	45	10,5	33	59	49,5	55	2,5	138	82	96	4
LLU	35	11	49,5	60	26	39	0	38	148	99	5
REA	9	13,5	3	85	11	77	19	20	88	94	6-7
LNAA	0	50	7,5	69	50	67	0	0	82	89	6-7
VeA	18	10	7,5	38	42,5	38	0,5	58	96	87	8
LMāA	6	13,5	3	10	50	52	0	0	154	97	9
RPIVA	68	14,5	12	46	38	47	0,5	0	68	88	10-12
LMūA	5	12	4,5	12	49	42	0,5	0	162	94	10-12
BA	50	18	4,5	33	26,5	31	0,5	0	120	94	10-12
LSPA	23	11	10,5	51	43,5	32	0	0	106	95	13
RA	63	12	10,5	36	39,5	65	2	0	48	86	14
ViA	30	11	4,5	27	34,5	66	0	20	80	83	15-17
LKuA	8	10,5	4,5	25	43	47	1	0	122	93	15-17
LJA	21	4,5	3	33	30,5	13	0,5	0	148	97	15-17
LiepU	39	13	13,5	49	21	38	0	0	72	93	18

Klasterizācijas rezultāti

Klas-teri	Augstskolas klasteros							
2	LU RTU RSU	Pārējās						
3	LU RTU	RSU	Pārējās					
4	LU RTU	RSU	DU	Pārējās				
5	LU RTU	RSU	DU	RPIVA RA ViA LiepU	Pārējās			
6	LU RTU	RSU	DU	RPIVA RA ViA LiepU	REA LNAA	Pārējās		
7	LU RTU	RSU	DU	RPIVA RA ViA LiepU	REA LNAA	LLU VeA BA LSPA	LMāA LKuA LMūA LJA	
8	LU RTU	RSU	DU	RPIVA RA ViA LiepU	REA	LLU VeA BA LSPA	LMāA LKuA LMūA LJA	LNAA

Ticamības analīze -1

Lai pārbaudītu veiktās klasterizācijas ticamību, tika izskaitļoti kvalitātes rādītāji - Randa un Huberta indeksi 2 līdz 8 klasteriem.

Izskaitļotās kopējās klasterizācijas kļūdas bija šādas:

2 klasteriem – 5,56 %;

3 klasteriem – 50%;

4 klasteriem – 55,6%;

5 klasteriem – 55,6%;

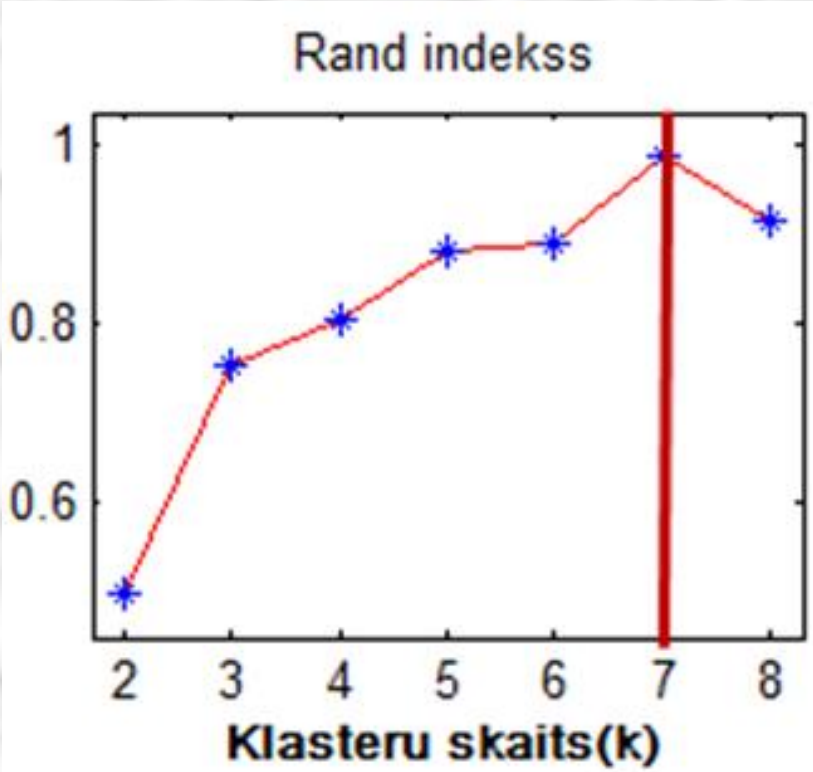
6 klasteriem – 38,89%;

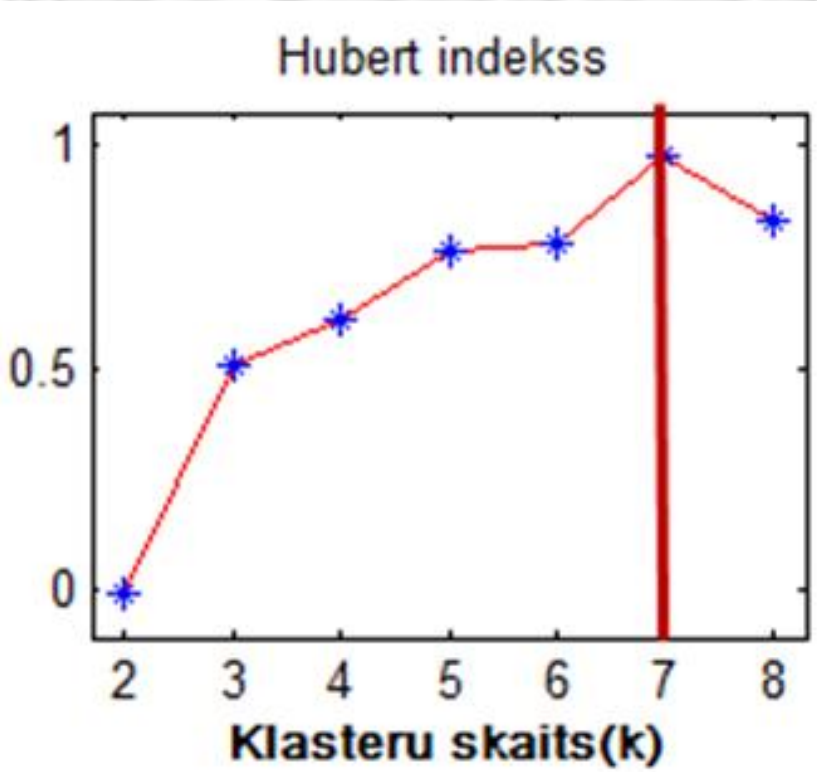
7 klasteriem – 33,33%;

8 klasteriem – 50%.

Ticamības analīze -2

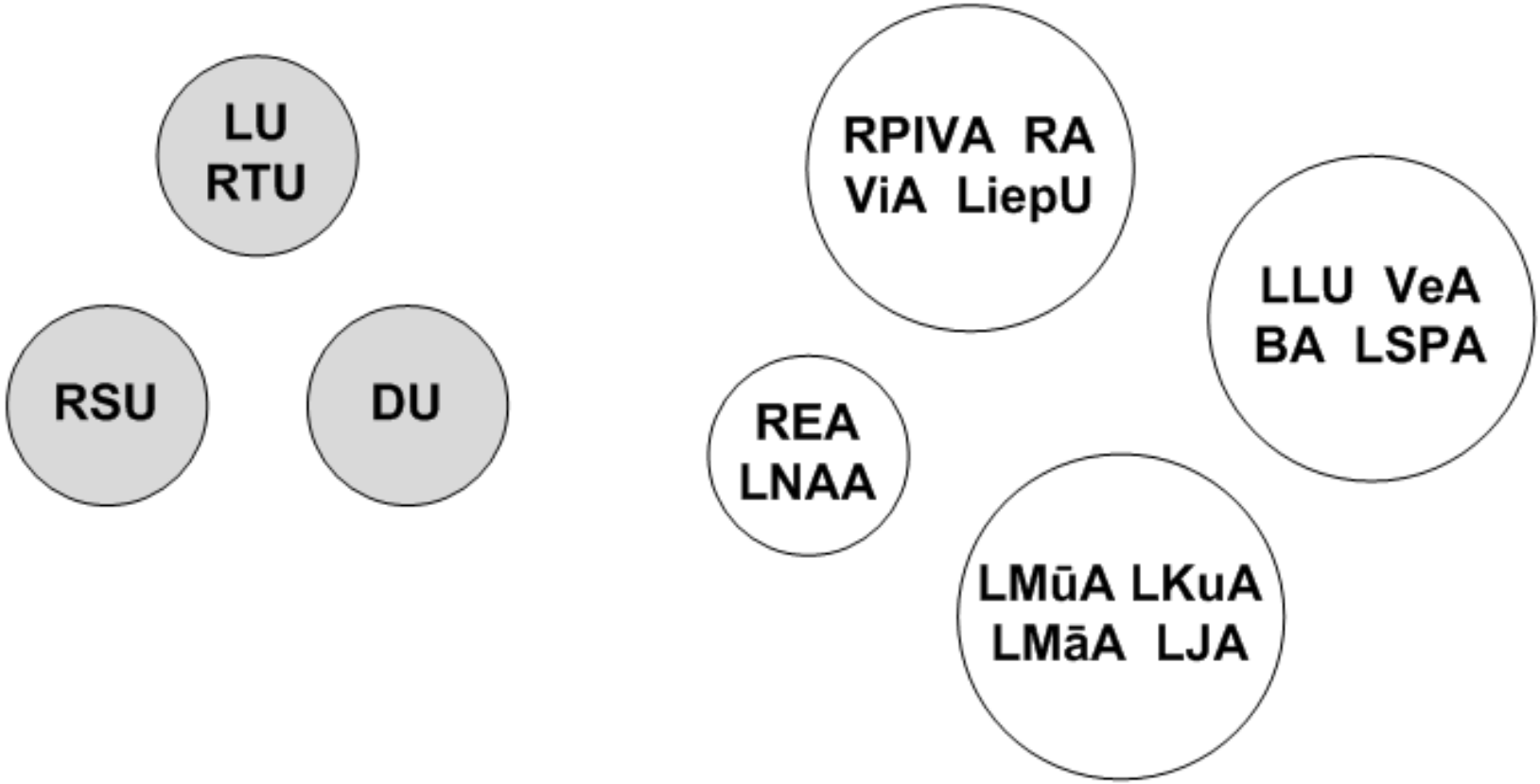
Randa un Huberta indeksi septiņu klasteru gadījumā:





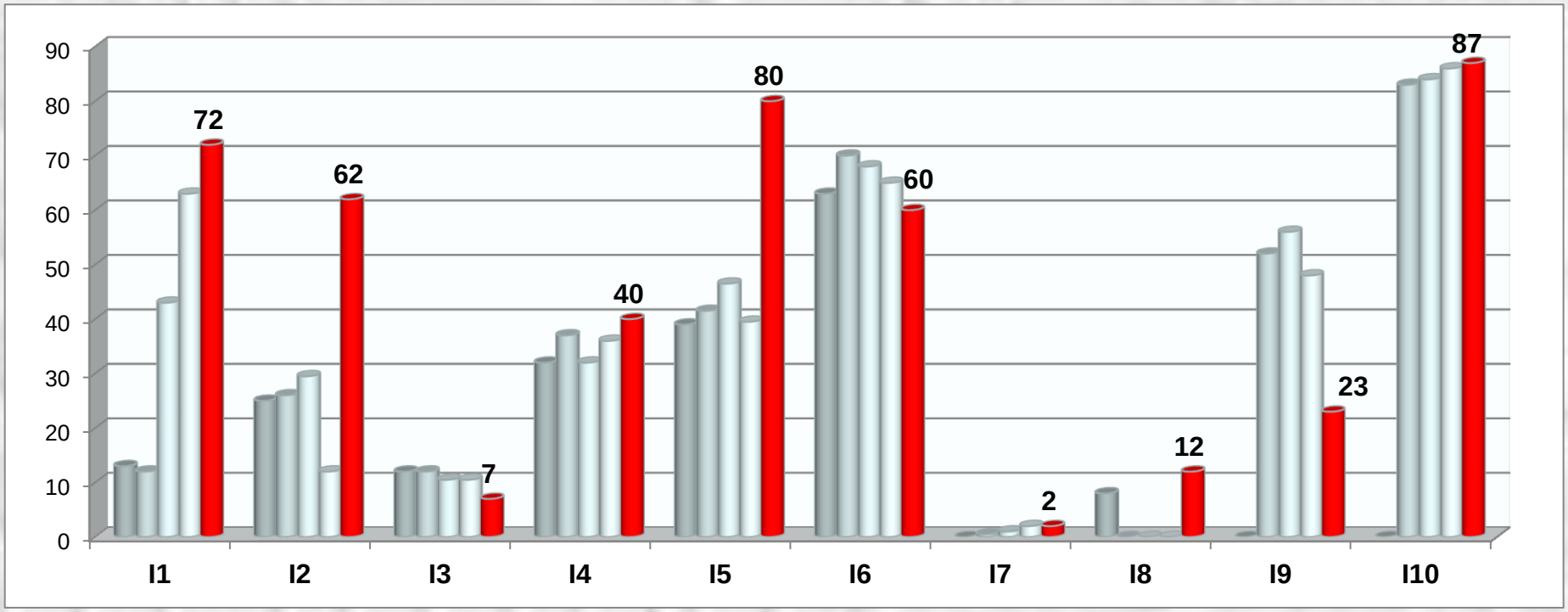
Rezultāts

Doto datu izlasi vislabāk raksturo 7 klasteru struktūra:



Rēzeknes Augstskola

Gads	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	Vieta
2008	13	25	12	32	39	63	0	8	-	-	7
2009	12	26	12	37	41.5	70	0.5	0	52	83	14
2010	43	29.5	10.5	32	46.5	68	1	0	56	84	13
2011	63	12	10.5	36	39.5	65	2	0	48	86	14
2012	72	62	7	40	80	60	2	12	23	87	11



PALDIES PAR UZMANĪBU !