Házi feladatok megoldása 12.

Különböző klasszifikációk összehasonlítása Centroid és Exacon módszerrel ROPstatban

Smahajcsik-Szabó Tamás, M9IJYM

1. A 10. óra 6. feladatában elmentett BIC-megoldás klaszterváltozójával végezz centroid-elemzést ROPstatban! Mely klaszterek vannak a legközelebb egymáshoz? Találsz itt összevonandó klasztereket?

A Bayes-féle információs kritérium nyomán G=3 és 10 között a G=3 megoldást fogadtam el az eredeti elemzés során.

MKA(G=3)	KL1	KL2	KL3
KL1	0.00	0.99	1.09
KL2	0.99	0.00	2.29
KL3	1.09	2.29	0.00

1. táblázat: A G=3 mclust() elemzés nyomán képzett klaszterek centroidjainak páronkénti távolságai.

Az ezen megoldás révén képzett klaszterek centroidjai kellőképp távol esnek (lásd a fenti táblázatban), így a háromklaszteres megoldás során nem szükséges további klaszterösszevonás.

2. Hasonlítsd össze a 8. óra 1. feladatának legjobb k-közép megoldását és a 10. óra 6. feladatában elmentett BIC-megoldás klaszterváltozóját a ROPstat centroid-elemzésével! Hány klaszterpár tekinthető igen közelinek?

KKA7	1	2	3	4	5	6	7
1	0.00	1.32	0.73	1.14	3.23	5.00	1.12
2	1.32	0.00	2.31	0.46	4.02	8.66	3.26
3	0.73	2.31	0.00	1.03	1.13	2.47	0.68
4	1.14	0.46	1.03	0.00	1.78	5.55	1.94
5	3.23	4.02	1.13	1.78	0.00	1.69	1.52
6	5.00	8.66	2.47	5.55	1.69	0.00	1.76
7	1.12	3.26	0.68	1.94	1.52	1.76	0.00

2. táblázat: A k=7 k-középpontú elemzés klasztereinek centroidjai közti páronkénti távolságok

A lenti, 3. táblázat alapján, a két megoldás klasztercentroidjainak távolságait tekintve, az MKA 1., 2. és 3. klasztereinek távolságai alacsonyak (<0.5) a KKA hasonló 1., 2. és 3. kódolású centroidjaival (rendre 0.04, 0.04 és 0.27). Néhány további hasonlóan közeli centroidpár jelenik meg: az MKA1 0.48 távolságot mutat a KKA3 klaszterrel, hasonlóképp az MKA2 és a KKA4 klaszterek közötti távolság is alacsony (0.23). Az MKA3 klaszter a KKA4-től 0.39, a KKA7-től pedig 0.48 távolságra fekszik.

MKA3	1	2	3	4	5	6	7
1	0.04	1.17	0.48	0.81	2.58	4.51	0.95
2	1.21	0.04	1.86	0.23	3.25	7.66	2.79
3	1.49	2.88	0.27	1.20	0.39	1.60	0.48

3. táblázat: A k=7 k-középpontú, és a G=3 modell alapú klaszterelemzések klasztereinek centroidjai közti páronkénti távolságok

Index	MKA	KKA	d
2	2	2	0.042
3	2	4	0.228
4	3	3	0.265
5	3	5	0.390
6	1	3	0.482
7	3	7	0.483
8	1	4	0.808
9	1	7	0.950
10	1	2	1.171
11	3	4	1.196
12	2	1	1.206
13	3	1	1.490
14	3	6	1.600
15	2	3	1.859
16	1	5	2.584
17	2	7	2.792
18	3	2	2.881
19	2	5	3.251
20	1	6	4.511
21	2	6	7.663

4. táblázat: A k=7 k-középpontú, és a G=3 modell alapú klaszterelemzések klasztercentroidjainak páronkénti távolságai

Az alábbi táblázatban megjelenő gyakoriságok alapján elmondható, hogy a KKA elemzés 5., 6. és 7. klaszterei alkotta almintát az MKA3 klaszter foglalja magában. A két elemzés adta struktúra jelentősen eltér: Köztük egyedük a 2. kódjelű klaszterek közös metszete képvisel egy nagyobb sűrűsödést (N=144).

MKA3	1	2	3	4	5	6	7	Total
1	75	1	19	17	0	0	0	112
2	0	144	0	85	0	0	0	229
3	2	2	31	34	32	25	33	159
Total	77	147	50	136	32	25	33	500

5. táblázat: A k=7 k-középpontú, és a G=3 modell alapú klaszterelemzések klasztereinek gyakorisági megoszlása

3. Hasonlítsd össze a 8. óra 1. feladatának legjobb k-közép megoldását és a 10. óra 6. feladatában elmentett BIC-megoldás klaszterváltozóját a ROPstat EXACON modulja segítségével! Mennyire tekinthető hasonlónak a két megoldás a Jaccard és a korrigált Rand index szerint?

A 0.20-nál nagyobb egyezési arányokról az alábbi táblázat tájékoztat.

KKA7-MKA3	1	2	3	Total
1	0.82	-	-	77
2	-	0.8	-	147
3	0.27	-	0.41	50
4	-	0.5	0.23	136
5	-	-	0.6	32
6	-	-	0.58	25
7	-	-	0.6	33
Total	112	229	159	500

6. táblázat: A k=7 k-középpontú, és a G=3 modell alapú megoldások klaszterei közti egyezési arányok

A két megoldás 1., 2. és 3. kódjelű klaszterei közötti egyezési arányok nagyobbak (0.82, 0.8 és a hármas klaszterek között 0.41). Korábbi, a centroidok távolsági elemzésénél már megfigyelt mintázat (az MKA3 klaszter sűríti magába a KKA 5., 6. és 7. klasztereit) itt is megjelenik, 0.58-0.6 egyezési arányokkal.

KKA7	1	2	3
1	T	A	A
2	A	T	A
3	.	A	T
4	(A)	T	.
5	A	A	T
6	(A)	A	T
7	A	A	T

7. táblázat: A k=7 k-középpontú, és a G=3 modell alapú megoldások klaszterei szerinti konfigurációk

A Jaccard-index (0.379) és a korrigált Rand-index (0.394) is a két struktúra jelentős eltérését jelzi. További eredményeket közöl a **Függelék** táblázata az ExaCon eredményt illetően.

4. Hasonlítsd össze a férfiak és a nők almintáját a 8. óra 1. feladatának legjobb k-közép megoldása segítségével! Melyik mintázatú klaszterben a legkisebb, illetve legnagyobb a nők aránya? Tudnál erre szakmai magyarázatot adni?

KL	Variable	Mean	Std.	Min	Max	Gender
1	PTelj	22.560	4.621	0	30	male
3	Pboldog	7.284	2.232	0	10	male
4	Pmagany	3.148	3.035	0	10	male
1	PTelj	22.880	4.046	6	30	female
3	Pboldog	7.524	2.002	0	10	female
4	Pmagany	3.176	3.084	0	10	female

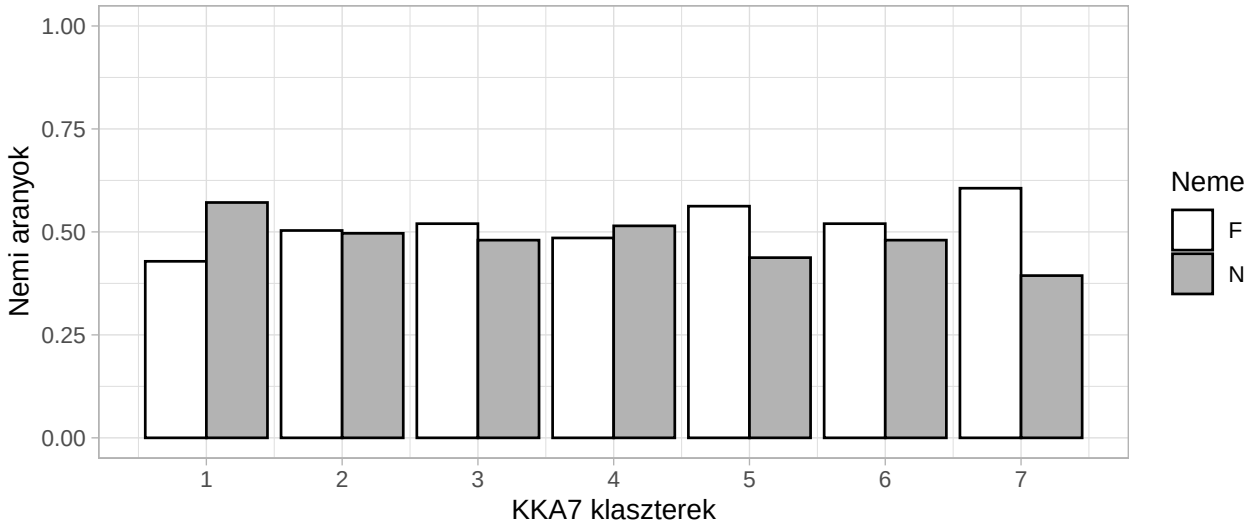
8. táblázat: Nemek szerinti leíró statisztikák a három változóra nézve

KKA7	1	2	3	4	5	6	7	Group
1	0.00	1.36	0.92	1.27	3.46	5.36	1.13	male
2	1.36	0.00	2.32	0.48	4.33	8.84	3.19	male
3	0.92	2.32	0.00	0.99	1.04	2.54	0.81	male
4	1.27	0.48	0.99	0.00	2.01	5.70	2.02	male
5	3.46	4.33	1.04	2.01	0.00	1.37	1.51	male
6	5.36	8.84	2.54	5.70	1.37	0.00	2.04	male
7	1.13	3.19	0.81	2.02	1.51	2.04	0.00	male
1	0.00	1.30	0.59	1.05	3.05	4.68	1.14	female
2	1.30	0.00	2.32	0.45	3.66	8.49	3.36	female
3	0.59	2.32	0.00	1.08	1.29	2.39	0.56	female
4	1.05	0.45	1.08	0.00	1.53	5.41	1.88	female
5	3.05	3.66	1.29	1.53	0.00	2.16	1.55	female
6	4.68	8.49	2.39	5.41	2.16	0.00	1.46	female
7	1.14	3.36	0.56	1.88	1.55	1.46	0.00	female
1	0.01	1.38	0.69	1.19	3.31	4.83	1.20	male_to_female
2	1.29	0.00	2.32	0.46	3.66	8.46	3.32	male_to_female
3	0.79	2.31	0.03	0.97	0.97	2.34	0.66	male_to_female
4	1.12	0.47	1.12	0.00	1.53	5.52	2.00	male_to_female
5	3.22	4.33	1.32	2.00	0.07	1.48	1.29	male_to_female
6	5.17	8.86	2.65	5.61	2.04	0.03	1.75	male_to_female
7	1.09	3.24	0.68	1.89	1.78	1.71	0.02	male_to_female

9. táblázat: A k=7 k-középpontú elemzés révén kapott 7 klaszter centroidjainak páronkénti távolsága nemi bontásban ("male" és "female" kódú sorok, illetve nemek közti összevetésben a "male to female" kódolással)

Mint a fenti, 9. táblázatból látható, az azonos kódolású klaszterek páronkénti távolságai nagyfokú közelséget jeleznek (egyedül az 5-5. klaszterek között emelkedik 0.07 szintre).

A KKA 7 klaszteres megoldásának nemi arányairól az alábbi ábra tájékoztat.



Nemek vonatkozásában a függetlenség esetén elvárt gyakoriságokhoz képest a KKA tekintetében nem mutatkoztak szignifikáns eltérések.

5. Melyik mintázatú klaszterben a legkisebb, illetve legnagyobb a 60 év felettiek aránya? Tudnál erre szakmai magyarázatot adni?

A k-középpontú elemzéssel nyert klaszterek közül a 4. és 5. klasztereknél a legmagasabb a 60 év felettiek aránya (34%), legalacsonyabb a 6. klaszternél (16%). Egyetlen megfigyelt cellagyakoriság sem különbözött szignifikánsan a függetlenség esetén elvárttól.

KKA7	n
1	0.260
2	0.299
3	0.200
4	0.346
5	0.344
6	0.160
7	0.273

10. táblázat: A k=7 k-középpontú elemzés révén kapott 7 klaszter szerinti életkori arány (60 év felettiek aránya)

MKA3	n
1	0.232
2	0.319
3	0.289

11. táblázat: A G=3 modell alapú elemzés révén kapott 3 klaszter szerinti életkori arány (60 év felettiek aránya)

Függelék

KKA7	stat.	1	2	3
1	Megfigyelt cellagyakorisag	75.0000	0.00	2.0000
1	Fuggetlenseg eseten vart cellagyakorisag	17.2000	35.30	24.5000
1	Khi-negyzet komponens	193.3700	35.27	20.6500
1	Hipergeometrikus valoszinuseg (egyoldalu)	0.0000	0.00	0.0000
1	Hipergeometrikus valoszinuseg (ketoldalu)	0.0000	0.00	0.0000
2	Megfigyelt cellagyakorisag	1.0000	144.00	2.0000
2	Fuggetlenseg eseten vart cellagyakorisag	32.9000	67.30	46.7000
2	Khi-negyzet komponens	30.9600	87.32	42.8300
2	Hipergeometrikus valoszinuseg (egyoldalu)	0.0000	0.00	0.0000
2	Hipergeometrikus valoszinuseg (ketoldalu)	0.0000	0.00	0.0000
3	Megfigyelt cellagyakorisag	19.0000	0.00	31.0000
3	Fuggetlenseg eseten vart cellagyakorisag	11.2000	22.90	15.9000
3	Khi-negyzet komponens	5.4300	22.90	14.3400
3	Hipergeometrikus valoszinuseg (egyoldalu)	0.0063	0.00	0.0000
3	Hipergeometrikus valoszinuseg (ketoldalu)	0.0112	0.00	0.0000
4	Megfigyelt cellagyakorisag	17.0000	85.00	34.0000
4	Fuggetlenseg eseten vart cellagyakorisag	30.5000	62.30	43.2000
4	Khi-negyzet komponens	5.9500	8.28	1.9800
4	Hipergeometrikus valoszinuseg (egyoldalu)	0.0006	0.00	0.0283
4	Hipergeometrikus valoszinuseg (ketoldalu)	0.0011	0.00	0.0521
5	Megfigyelt cellagyakorisag	0.0000	0.00	32.0000
5	Fuggetlenseg eseten vart cellagyakorisag	7.2000	14.70	10.2000
5	Khi-negyzet komponens	7.1700	14.66	46.8000
5	Hipergeometrikus valoszinuseg (egyoldalu)	0.0002	0.00	0.0000
5	Hipergeometrikus valoszinuseg (ketoldalu)	0.0003	0.00	0.0000
6	Megfigyelt cellagyakorisag	0.0000	0.00	25.0000
6	Fuggetlenseg eseten vart cellagyakorisag	5.6000	11.40	8.0000
6	Khi-negyzet komponens	5.6000	11.45	36.5700
6	Hipergeometrikus valoszinuseg (egyoldalu)	0.0015	0.00	0.0000
6	Hipergeometrikus valoszinuseg (ketoldalu)	0.0023	0.00	0.0000
7	Megfigyelt cellagyakorisag	0.0000	0.00	33.0000
7	Fuggetlenseg eseten vart cellagyakorisag	7.4000	15.10	10.5000
7	Khi-negyzet komponens	7.3900	15.11	48.2700
7	Hipergeometrikus valoszinuseg (egyoldalu)	0.0002	0.00	0.0000
7	Hipergeometrikus valoszinuseg (ketoldalu)	0.0003	0.00	0.0000

12. táblázat: Khi-négyzet próba eredmények az MKA és KKA klaszterbesorolásai szerint