NOMCLUST 2.0:

BALÍČEK PRO SHLUKOVÁNÍ OBJEKTŮ CHARAKTERIZOVANÝCH KATEGORIÁLNÍMI PROMĚNNÝMI

Zdeněk Šulc
Vysoká škola ekonomická v Praze

Robust 2018

OBSAH

- současná verze balíčku nomclust
- nová evaluační kritéria
- výpočetní optimalizace

SOUČASNÁ VERZE BALÍČKU NOMCLUST

- shlukování objektů charakterizovaných kategoriálními proměnnými pomocí:
  - koeficientu prosté shody
  - měr podobnosti pro binární data (vyžadují binární transformaci dat)

SOUČASNÁ VERZE BALÍČKU NOMCLUST

- shlukování objektů charakterizovaných kategoriálními proměnnými pomocí:
  - koeficientu prosté shody
  - měr podobnosti pro binární data (vyžadují binární transformaci dat)
  - měr podobnosti pro nominální data

SOUČASNÁ VERZE BALÍČKU NOMCLUST

- představen v (Šulc a Řezanková, 2015)
- 3 metody hierarchické shlukové analýzy (complete, average, single)
- 13 měr podobnosti
- 6 evaluačních kritérií výsledných shluků
- k dispozici v repozitáři CRAN pod jménem „nomclust“

SOUČASNÁ VERZE BALÍČKU NOMCLUST

- 13 měr podobnosti (Boriah a kol., 2008), (Morlini a Zani, 2012), (Šulc, 2016):

Eskin (eskin)	Lin 1 (lin1)
Goodall 1 (good1)	Morlini and Zani (morlini)
Goodall 2 (good2)	Occurrence Freq. (of)
Goodall 3 (good3)	Simple Matching Coef. (sm)
Goodall 4 (good4)	Variable Entropy (ve)
Inverse Occurrence Freq. (iof)	Variable Mutability (vm)
Lin (lin)	

SOUČASNÁ VERZE BALÍČKU NOMCLUST

- 4 kritéria pro ohodnocení vnitroshlukové variability
- 2 kritéria pro určení optimálního počtu shluků

Koeficient vnitroshlukové mutability (WCM)	vnitroshluková variabilita
Koeficient vnitroshlukové entropie (WCE)	
Pseudo tau koeficient (PSTau)	
Pseudo koeficient neurčitosti (PSU)	
Pseudo F koeficient založený na mutabilitě (PSFM)	optimální počet shluků
Pseudo F koeficient založený na entropii (PSFE)	

SOUČASNÁ VERZE BALÍČKU NOMCLUST

- 4 typická použití R balíčku:
  - výpočet matice nepodobností pro určitou míru podobnosti
  - kompletní hierarchické shlukování
  - výpočet hodnotících kritérií u souborů s již vypočítanými příslušnostmi objektů pomocí jiné metody
  - využití již vypočtené matice nepodobností pro shlukování

NOVÁ EVALUAČNÍ KRITÉRIA

- 3 nová evaluační kritéria pro určení optimálního počtu shluků
  - BIC založené na entropii (BIC1)
  - BIC založené na mutabilitě (BIC2)
  - BK index

NOVÁ EVALUAČNÍ KRITÉRIA

- BIC založené na entropii (BIC1)

$$BIC1(k) = 2\sum_{g=1}^{k} n_g \sum_{c=1}^{m} H_{gc} + k\sum_{c=1}^{m}(K_c - 1)\ln(n)$$

$$H_{gc} = -\sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\ln\frac{n_{gcu}}{n_g}\right)$$

k – počet shluků, n_g – počet objektů v g-tém shluku, n – počet objektů,
H_{gc} – entropie c-té proměnné v g-tém shluku, K_c – počet kategorií c-té proměnné.

NOVÁ EVALUAČNÍ KRITÉRIA

- BIC založené na mutabilitě (BIC2)

$$BIC2(k) = 2\sum_{g=1}^{k} n_g \sum_{c=1}^{m} G_{gc} + k\sum_{c=1}^{m}(K_c - 1)\ln(n)$$

$$G_{gc} = 1 - \sum_{u=1}^{K_c}\left(\frac{n_{gcu}}{n_g}\right)^2$$

k – počet shluků, n_g – počet objektů v g-tém shluku, n – počet objektů,
G_{gc} – mutabilita c-té proměnné v g-tém shluku, K_c – počet kategorií c-té proměnné.

NOVÁ EVALUAČNÍ KRITÉRIA

- výpočetní optimalizace po vzoru SPSS (Bacher a kol., 2004):

$$R(k) = \frac{d_{k-1}}{d_k} \qquad d_k = \sum_{h=1}^{k-1}\sum_{c=1}^{m} H_{gc} - \sum_{g=1}^{k}\sum_{c=1}^{m} H_{gc} \qquad r = \frac{R(k_{(1)})}{R(k_{(2)})}$$

hodnota r	optimální počet shluků
$r > 1{,}15$	$k_{(1)}$
$r \leq 1{,}15$	$\max\{k_{(1)}, k_{(2)}\}$

NOVÁ EVALUAČNÍ KRITÉRIA

- BK index

$$BK(k) = \Delta^2 I(k) = (I(k-1) - I(k)) - (I(k) - I(k+1))$$

$$I(k) = H_E(k) - H_E(k+1)$$

$$H_E(k) = \sum_{g=1}^{k} \frac{n_g}{n} \sum_{c=1}^{m} \frac{H_{gc}}{\ln K_c}$$

k – počet shluků, n_g – počet objektů v g-tém shluku, n – počet objektů,
H_{gc} – entropie c-té proměnné v g-tém shluku, K_c – počet kategorií c-té proměnné.

NOVÁ EVALUAČNÍ KRITÉRIA

```
R> clu_eval <- data$eval
```

	cluster	WCM	WCE	PSTau	PSU	PSFM	PSFE
1	1	0.9666	0.9600	NA	NA	NA	NA
2	2	0.7330	0.7010	0.1987	0.2084	4.4646	4.7382
3	3	0.6127	0.5789	0.3365	0.3607	4.3106	4.7949
4	4	0.4546	0.4066	0.5005	0.5491	5.3446	6.4961
5	5	0.4136	0.3658	0.5373	0.5807	4.3551	5.1943
6	6	0.3600	0.3085	0.6004	0.6514	4.2074	5.2327

NOVÁ EVALUAČNÍ KRITÉRIA

```
R> clu_eval <- data$eval
```

	cluster	WCM	WCE	BIC1	BIC2	PSFM	PSFE	BK
1	1	0.967	0.960	NA	NA	NA	NA	NA
2	2	0.733	0.701	54.59	77.74	4.465	4.738	0.837
3	3	0.613	0.579	89.50	120.51	4.311	4.795	-0.470
4	4	0.455	0.407	125.24	163.74	5.345	6.496	0.305
5	5	0.414	0.366	162.27	208.41	4.355	5.194	0.067
6	6	0.360	0.309	196.69	251.03	4.207	5.233	-0.187
7	opt	NA	NA	2	5	3	3	2

VÝPOČETNÍ OPTIMALIZACE

- výpočet matice nepodobností je výpočetně nejnáročnější část hierarchického shlukování
  - výpočetní náročnost roste čtvercem počtu pozorování
- velké množství shodných objektů v kategoriálních datech
  - časté opakování stejných výpočtů

VÝPOČETNÍ OPTIMALIZACE

- výpočet matice nepodobností je výpočetně nejnáročnější část hierarchického shlukování
  - výpočetní náročnost roste čtvercem počtu pozorování
- velké množství shodných objektů v kategoriálních datech
  - časté opakování stejných výpočtů
  - možnost redukce rozměru úlohy

VÝPOČETNÍ OPTIMALIZACE

návrh algoritmu pro zrychlení výpočtu:

1. odstranění duplicitních objektů a vytvoření vektoru vah
2. výpočet matice nepodobností pouze pro jedinečné objekty
3. pomocí vektoru vah je zrekonstruována původní matice vzdáleností
4. provedení aglomerativní shlukové analýzy

VÝPOČETNÍ OPTIMALIZACE

překódování hodnot

v1	v2	v3	v4
1	2	1	5
3	1	2	4
2	1	1	3
1	2	1	5
3	1	2	4
2	2	1	3
2	1	1	3
1	1	1	5
1	2	1	5
3	2	2	4

v1	v2	v3	v4
1	1	1	1
2	2	2	2
1	1	1	1
3	2	1	3
1	1	1	1
2	2	2	2
3	1	1	3
3	2	1	3
1	2	1	1
1	1	1	1
2	1	2	2

VÝPOČETNÍ OPTIMALIZACE

seřazení hodnot

v1	v2	v3	v4
1	1	1	1
2	2	2	2
1	1	1	1
3	2	1	3
1	1	1	1
2	2	2	2
3	1	1	3
3	2	1	3
1	2	1	1
1	1	1	1
2	1	2	2

v1	v2	v3	v4
1	1	1	1
1	1	1	1
1	1	1	1
1	2	1	1
2	1	2	2
2	2	2	2
2	2	2	2
3	1	1	3
3	2	1	3
3	2	1	3

VÝPOČETNÍ OPTIMALIZACE

redukce rozměru dat

v1	v2	v3	v4
1	1	1	1
1	1	1	1
1	1	1	1
1	1	1	1
1	2	1	1
2	1	2	2
2	2	2	2
2	2	2	2
2	2	2	2
3	1	1	3
3	2	1	3
3	2	1	3

v1	v2	v3	v4	ni
1	1	1	1	4
1	2	1	1	1
2	1	2	2	1
2	2	2	2	2
3	1	1	3	1
3	2	1	3	2

VÝPOČETNÍ OPTIMALIZACE

rekonstrukce původní matice vzdáleností

0.000	0.091	0.211	0.347
0.091	0.000	0.347	0.211
0.211	0.347	0.000	0.091
0.347	0.211	0.091	0.000

vektor vah

4	1	1	2

0.000	0.000	0.000	0.000	0.091	0.211	0.347	0.347
0.000	0.000	0.000	0.000	0.091	0.211	0.347	0.347
0.000	0.000	0.000	0.000	0.091	0.211	0.347	0.347
0.000	0.000	0.000	0.000	0.091	0.211	0.347	0.347
0.091	0.091	0.091	0.091	0.000	0.347	0.091	0.091
0.211	0.211	0.211	0.211	0.347	0.000	0.091	0.091
0.347	0.347	0.347	0.347	0.211	0.091	0.000	0.000
0.347	0.347	0.347	0.347	0.211	0.091	0.000	0.000

ZÁVĚR

- balíček *nomclust* nabízí míry podobnosti, které se nenachází v ostatních R balíčcích ani v komerčním softwaru
- v nové verzi byla zachována původní funkcionalita
  - přidána nová hodnotící kritéria pro určení optimálního počtu shluků
  - provedené optimalizace výrazně zrychlují výpočet

REFERENCE

Bacher, J., Wenzig, K. a Vogler, M. 2004. SPSS TwoStep Cluster - a first evaluation. Lehrstuhl für Soziologie, Nürnberg.

Boriah, S., Chandola, V. a Kumar, V. 2008. Similarity measures for categorical data: A comparative evaluation. In Proceedings of the 8th SIAM International Conference on Data Mining, SIAM, pp. 243-254.

Morlini, I., Zani, S. 2012. A new class of weighted similarity indices using polytomous variables. In Journal of Classification, 29 (2), pp. 199-226.

Šulc, Z. a Řezanková, H. 2015. nomclust: An R package for hierarchical clustering of objects characterized by nominal variables. In Proceedings of the 9th International Days of Statistics and Economics. Melandrium, Slaný, pp. 1581-1590.

Šulc, Z. 2016. Similarity measures for nominal data in hierarchical clustering. Disertační práce, Vysoká škola ekonomická v Praze.