Metodologie hodnocení měr podobnosti pro kategoriální data na generovaných datových souborech

Zdeněk Šulc

Vysoká škola ekonomická v Praze

Robust 2016

Obsah

- Současné přístupy hodnocení měr podobnosti v hierarchické shlukové analýze (HCA)
- Nově navržená metodologie
- Ukázka praktického použití
- Závěr

Současné přístupy hodnocení měr podobnosti v HCA

- na několika reálných souborech, většinou ze známých repozitářů
- za použití nízkého počtu generovaných souborů
- externí kritéria hodnocení shlukování
  - shlukování vs. klasifikace

Nově navržená metodologie

- hodnocení měr podobnosti v HCA:
  - na základě velkého množství generovaných souborů
  - interní kritéria hodnocení shlukování (PSFE, PSFM)
  - založena na pořadových skórech koeficientů

Nově navržená metodologie

- koeficient PSFE

$$PSFE(k) = \frac{(n-k)\left[nWCE(1) - nWCE(k)\right]}{(k-1)nWCE(k)}$$

- koeficient nWCE

$$nWCE(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \left(- \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} ln \frac{n_{gcu}}{n_g} \right) \right)$$

n - počet objektů, *k* - počet shluků, *m* - počet proměnných, *K* - počet kategorií

Nově navržená metodologie

- koeficient PSFM

$$PSFM(k) = \frac{(n-k)\left[nWCM(1) - nWCM(k)\right]}{(k-1)nWCM(k)}$$

- koeficient nWCM

$$nWCM(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \left(1 - \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g}\right)^2\right)$$

n - počet objektů, *k* - počet shluků, *m* - počet proměnných, *K* - počet kategorií

Nově navržená metodologie

1. míry podobnosti jsou seřazeny sestupně podle hodnot PSFE a PSFM koeficientů pro každý soubor a počet shluků

measure	clu	psfe	r_psfe	psfm	r_psfm
ES	3	12.84	3	13.33	4
IOF	3	14.51	1.5	15.11	1.5
LIN	3	14.51	1.5	15.11	1.5
SM	3	10.77	4	12.09	3
OF	3	11.39	5	13.41	5

Nově navržená metodologie

2. pro každou míru podobnosti jsou spočteny průměrné hodnoty pořadí a směrodatné odchylky

measure	avg_psfm	avg_psfe	sd_psfm	sd_psfe
ES	2.1	2.0	1.3	1.2
IOF	1.8	1.8	1.2	1.2
LIN	3.1	3.4	1.4	1.3
SM	3.6	3.4	1.3	1.3
OF	3.9	4.1	1.2	1.3

Nově navržená metodologie

- výsledky zobrazeny ve formě přehledné tabulky
- třídění výstupů podle přednastavených parametrů
  - počet proměnných, počet shluků, komplikovanost souboru
- využívá R balíček *nomclust*
- pro generování doporučeny funkce *gen_object,* využívající balíček *clusterGeneration*

Ukázka praktického použití

- cílem je prozkoumání kvality 13 měr podobnosti pro nominální data v HCA (úplné spojení)

Eskin (ES)	Lin 1 (L1)
Goodall 1 (G1)	Morlini and Zani (MZ)
Goodall 2 (G2)	Occurrence Freq. (OF)
Goodall 3 (G3)	Simple Matching Coef. (SM)
Goodall 4 (G4)	Variable Entropy (VE)
Inverse Occurrence Freq. (IOF)	Variable Mutability (VM)
Lin (LIN)	

Ukázka praktického použití

- 60 generovaných souborů (funkce *gen_object*)
  - počty proměnných: 4, 6, 8, 10
    - komplikovanost souborů: lehká, střední, vysoká (easy, medium, hard)
    - počet shluků: 3
    - počet replikací: 5

Ukázka praktického použití

- na všech souborech provedena HCA s každou z 13 měr podobnosti
- vypočítány shlukové příslušnosti pro 2 až 6 shluků
- celkem 3900 výstupů HCA
  - 60 souborů x 13 měr podobnosti x 5 shluků

Celkové výsledky

	ES	G1	G2	G3	G4	IOF	LIN	L1	MZ	OF	SM	VE	VM
mean	5.0	8.3	7.8	8.5	8.3	5.2	5.2	5.8	9.3	6.0	8.9	6.7	6.0
sd	2.4	2.5	2.8	2.8	2.4	2.4	2.8	3.0	2.6	2.8	2.8	2.7	2.7

- nejlepší výsledky podává míra ES
- nejhůře se umístila míra MZ následovaná koeficientem prosté shody

Výsledky podle počtu proměnných

	ES	G1	G2	G3	G4	IOF	LIN	L1	MZ	OF	SM	VE	VM
4 var	4.2	10.0	7.7	10.0	8.9	4.1	6.5	5.7	8.4	9.7	8.4	3.7	3.7
6 var	3.6	8.0	8.0	8.7	7.3	5.6	5.8	5.6	10.1	7.2	9.9	6.0	5.3
8 var	5.4	8.6	7.4	7.6	8.3	4.8	4.3	6.2	10.3	4.5	8.3	8.2	7.2
10 var	6.8	6.7	7.9	7.6	8.6	6.4	4.2	5.8	8.3	2.8	8.9	8.9	8.0

- IOF, ES mají celkově dobré výsledky
- VE a VM upřednostňují menší počet proměnných
- OF, LIN podává lepší výsledky při vyšším počtu proměnných

Závěr

- jednoduchý přístup hodnocení měr podobnosti pro kategoriální proměnné v HCA
- možnost třídění výsledků podle různých kritérií
- schopnost přehledně ohodnotit velké množství generovaných souborů
- využití interní evaluace vhodné pro shlukovou analýzu