POROVNÁNÍ NOVÝCH PŘÍSTUPŮ V OBLASTI MĚR PODOBNOSTI PRO KATEGORIÁLNÍ DATA

Zdeněk Šulc
Vysoká škola ekonomická v Praze

Robust 2014

OBSAH

- Současné možnosti shlukování kategoriálních dat
- Nedávno představené míry podobnosti
- Experiment
- Závěry

SOUČASNÉ MOŽNOSTI

SHLUKOVÁNÍ KATEGORIÁLNÍCH DAT

OVERLAP

koeficient prosté shody

1. úroveň

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ jestliže } x_{ic} = x_{jc} \\ 0 \text{ jinak} \end{cases}$$

2. úroveň

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}$$

míra vzdálenosti

$$D_{ij} = 1 - S_{ij}$$

OVERLAP

	V1	V2	V3	V4	V5
Xi	a	d	a	a	c
Xj	b	d	d	b	c
$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ jestliže } x_{ic} = x_{jc} \\ 0 \text{ jinak} \end{cases}$	0	1	0	0	1

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m} = \frac{2}{5} = 0.4$$

$$D_{ij} = 1 - S_{ij} = 1 - 0.4 = 0.6$$

DVOUKROKOVÁ SHLUKOVÁ ANALÝZA

- 1. krok - vytvoření pomocných shluků
- 2. krok - aplikace shlukovacích algoritmů na pomocné shluky
- spojitá i kategoriální data
- závislá na počátečním pořadí objektů

NEDÁVNO PŘEDSTAVENÉ MÍRY PODOBNOSTI

ESKIN

větší váhy u proměnných s více kategoriemi

1. úroveň

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ jestliže } x_{ic} = x_{jc} \\ \dfrac{n_c^2}{n_c^2 + 2} \text{ jinak} \end{cases}$$

2. úroveň

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}$$

míra vzdálenosti

$$D_{ij} = \frac{1}{S_{ij}} - 1$$

OF

vyšší váhy u četných kategorií

1. úroveň

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ jestliže } x_{ic} = x_{jc} \\ \dfrac{1}{1 + \ln \dfrac{n}{f(x_{ic})} \cdot \ln \dfrac{n}{f(x_{jc})}} \text{ jinak} \end{cases}$$

2. úroveň

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}$$

míra vzdálenosti

$$D_{ij} = \frac{1}{S_{ij}} - 1$$

IOF

vyšší váhy u vzácných kategorií

1. úroveň

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ jestliže } x_{ic} = x_{jc} \\ \dfrac{1}{1 + \ln f(x_{ic}) \cdot \ln f(x_{jc})} \text{ jinak} \end{cases}$$

2. úroveň

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}$$

míra vzdálenosti

$$D_{ij} = \frac{1}{S_{ij}} - 1$$

LIN

shoda: vyšší váhy u četných kategorií
neshoda: nižší váhy u vzácných kategorií

1. úroveň

$$S_c(x_{ic}, x_{jc}) = \begin{cases} 2 \cdot \ln p(x_{ic}) \text{ jestliže } x_{ic} = x_{jc} \\ 2 \cdot \ln(p(x_{ic}) + p(x_{jc})) \text{ jinak} \end{cases}$$

2. úroveň

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m} (\ln p(x_{ic}) + \ln p(x_{jc}))}$$

míra vzdálenosti

$$D_{ij} = \frac{1}{S_{ij}} - 1$$

EXPERIMENT

DATA

- 4 datové soubory z UCI data depository
  http://archive.ics.uci.edu/ml/datasets.html

	Car	Hayes Roth	Breast Cancer	Post Operative
počet objektů	132	1728	683	90
počet proměnných	5	6	9	7
min. # kategorií	3	3	9	2
max. # kategorií	4	4	10	3

HODNOTÍCÍ KRITÉRIA SHLUKŮ

- normalizovaný Giniho koeficient

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

$$G'(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \frac{K_c}{K_c - 1} G_{gc}$$

- normalizovaná entropie

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

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

HODNOTÍCÍ KRITÉRIA SHLUKŮ

- Pseudo F index založený Giniho koeficientu

$$I_{\text{PSF-tau}}(k) = \frac{(n-k)(G(1)-G(k))}{(k-1)G(k)}$$

- Pseudo F index založený na entropii

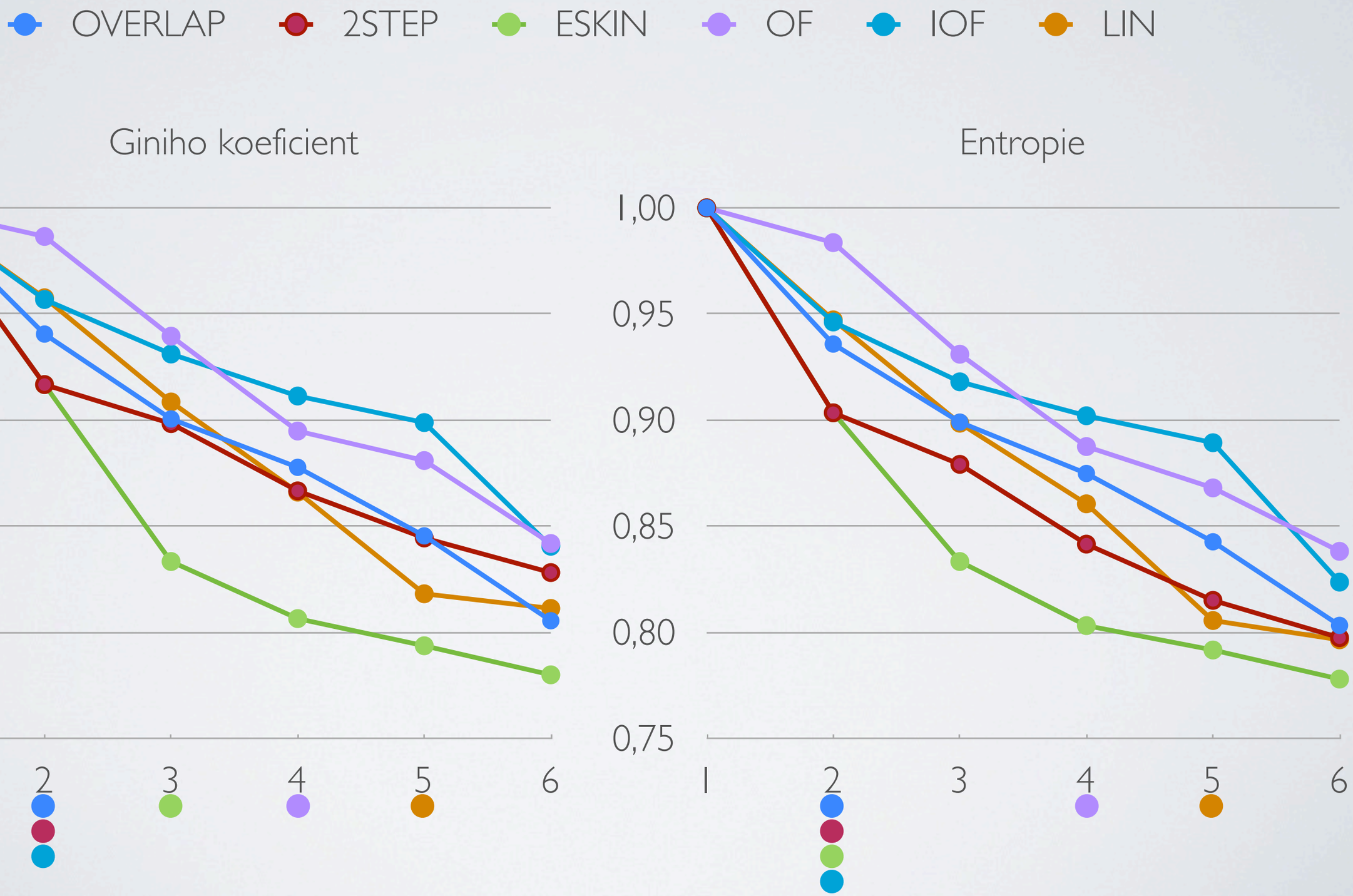
$$I_{\text{PSF-U}}(k) = \frac{(n-k)(H(1)-H(k))}{(k-1)H(k)}$$

HODNOCENÍ SHLUKŮ

CAR - OVERLAP

počet shluků	1	2	3	4	5	6
Gnorm	1,0000	0,9405	0,9005	0,8777	0,8454	0,8054
Hnorm	1,0000	0,9358	0,8990	0,8748	0,8426	0,8032
PSF-Tau	-	102,4332	89,2050	74,9161	73,7406	77,5811
PSF-U	-	103,8027	84,7088	71,9166	70,4517	73,3204

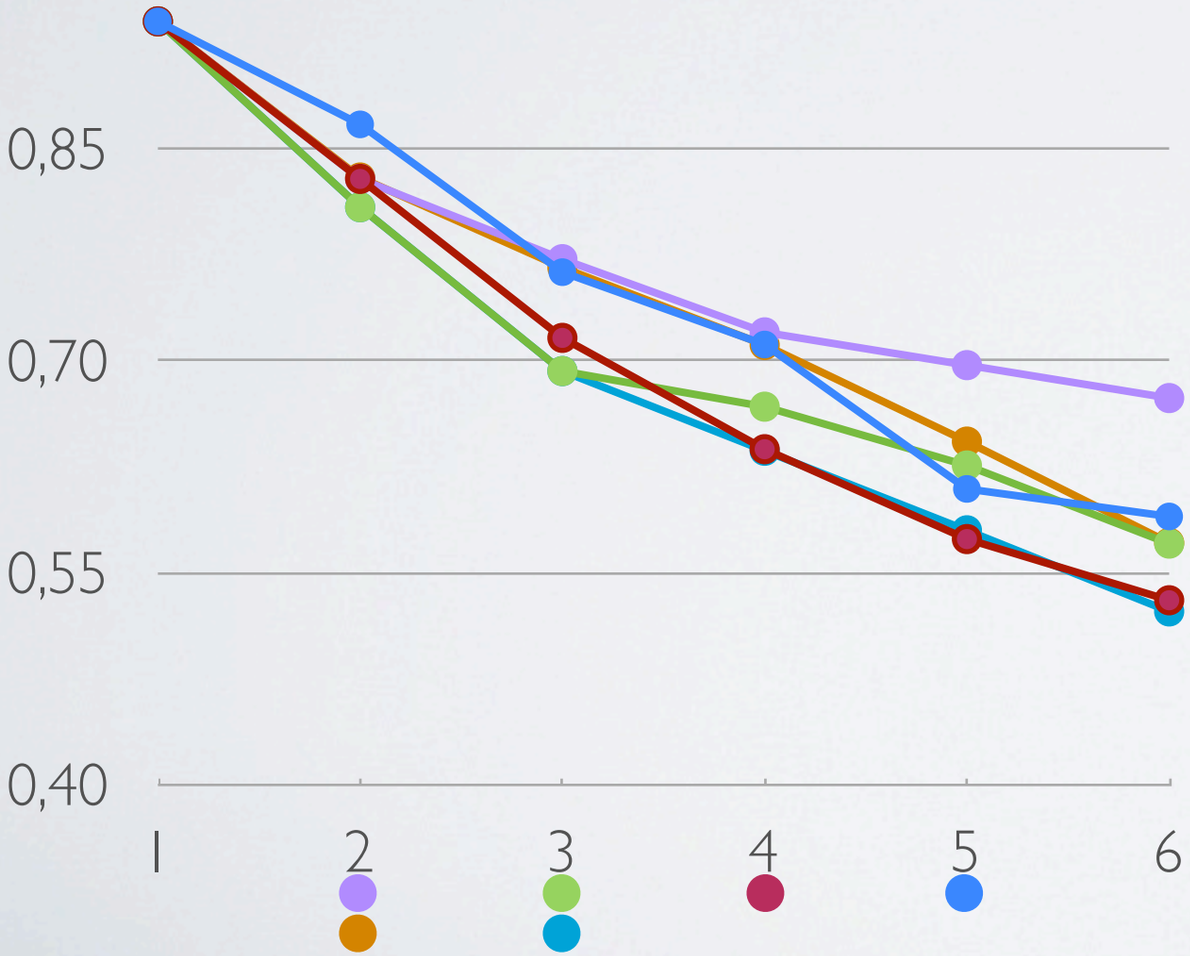
CAR


OVERLAP
2STEP
ESKIN
OF
IOF
LIN
Giniho koeficient
1,00
0,95
0,90
0,85
0,80
0,75
1
2
3
4
5
6
Entropie
1,00
0,95
0,90
0,85
0,80
0,75
1
2
3
4
5
6

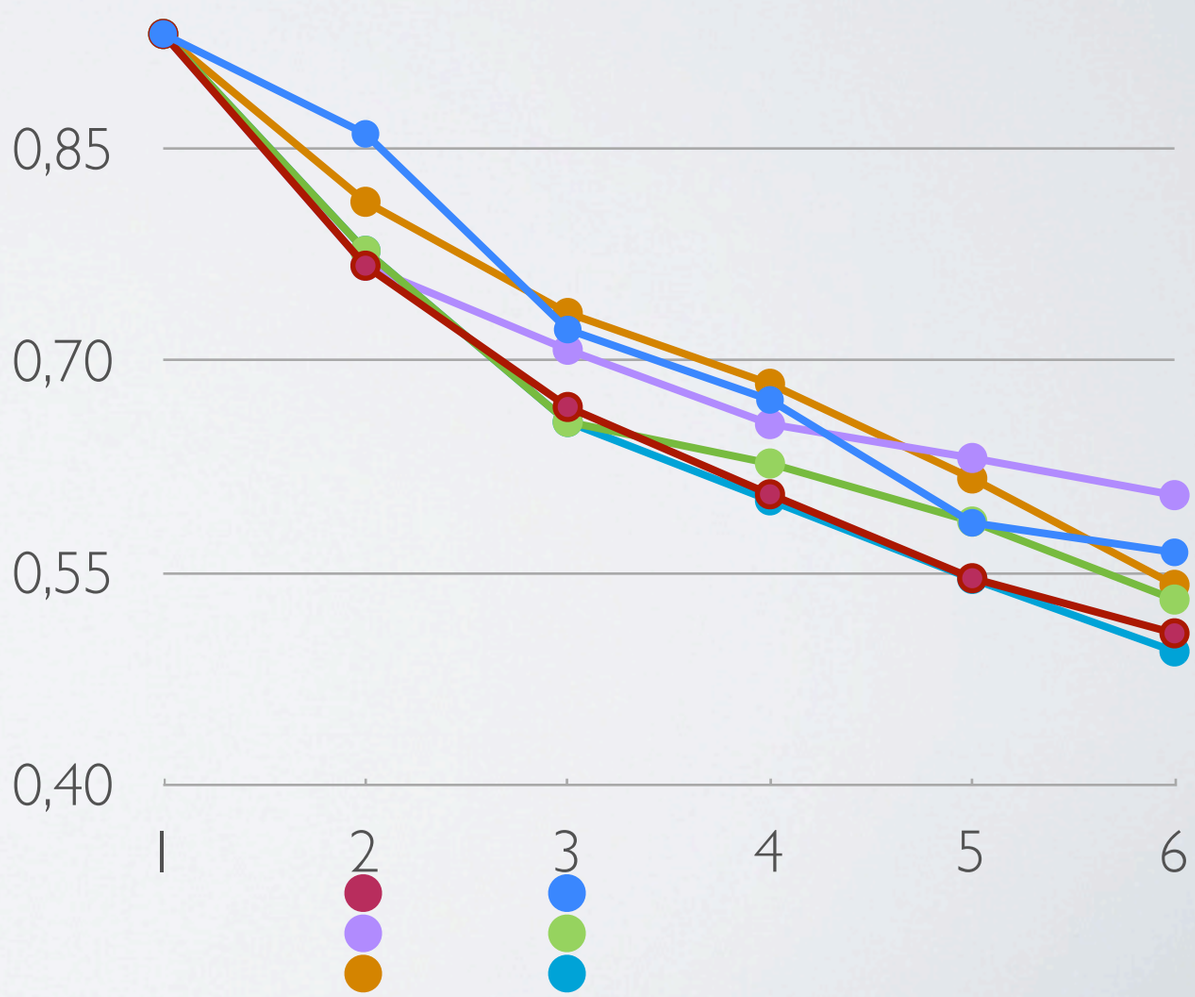
HAYES ROTH

OVERLAP 2STEP ESKIN OF IOF LIN

Giniho koeficient

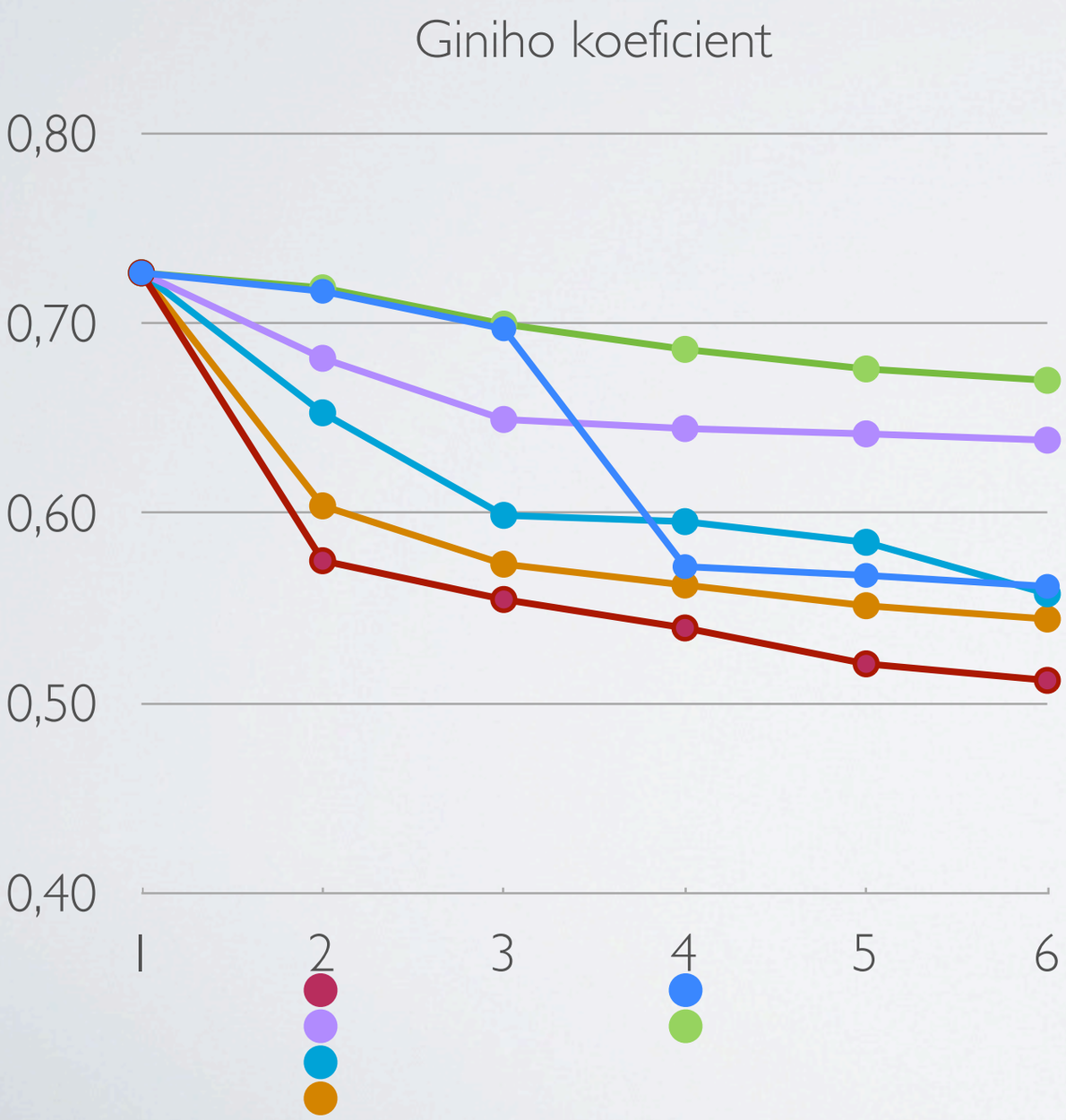


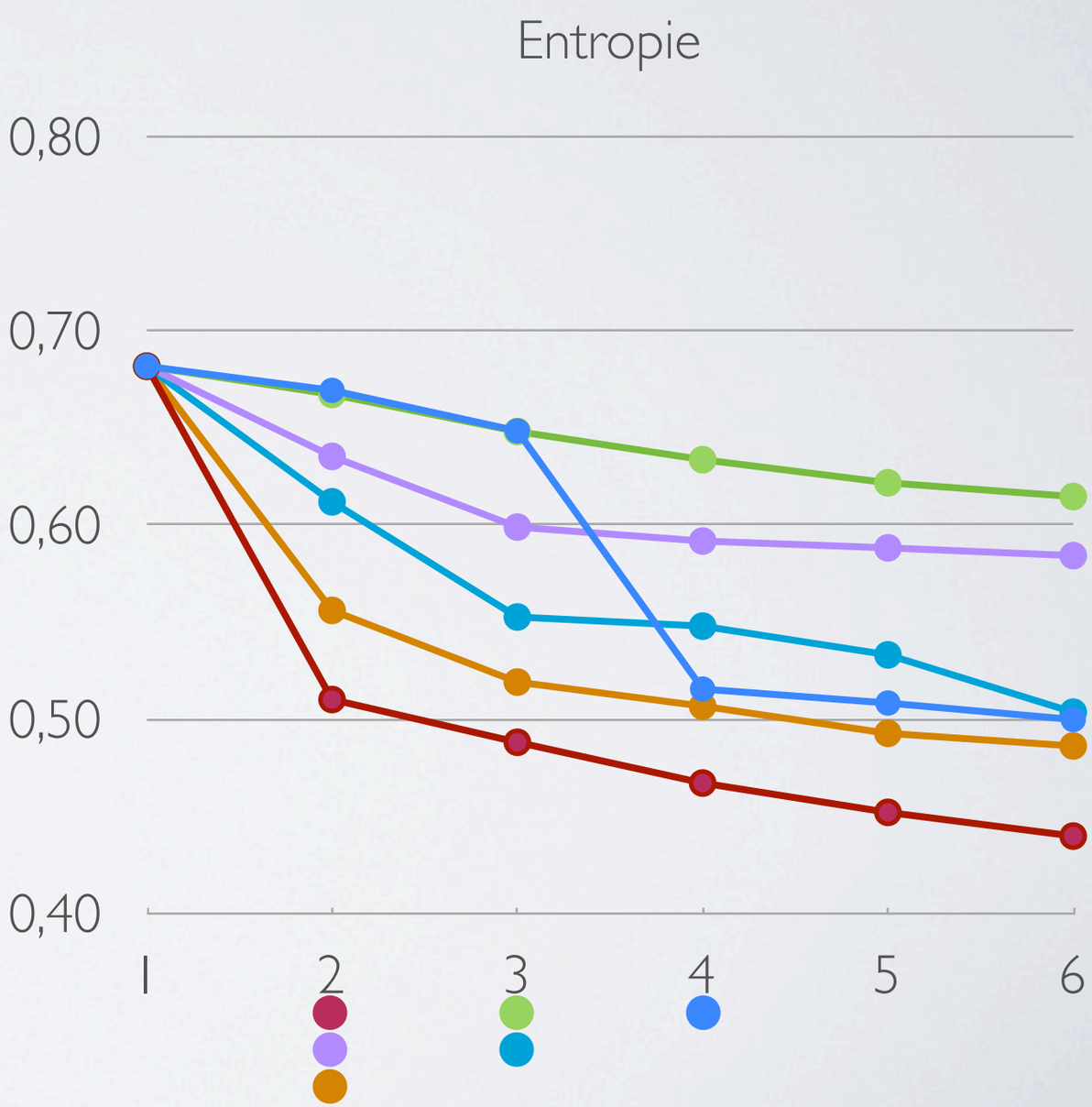
Entropie



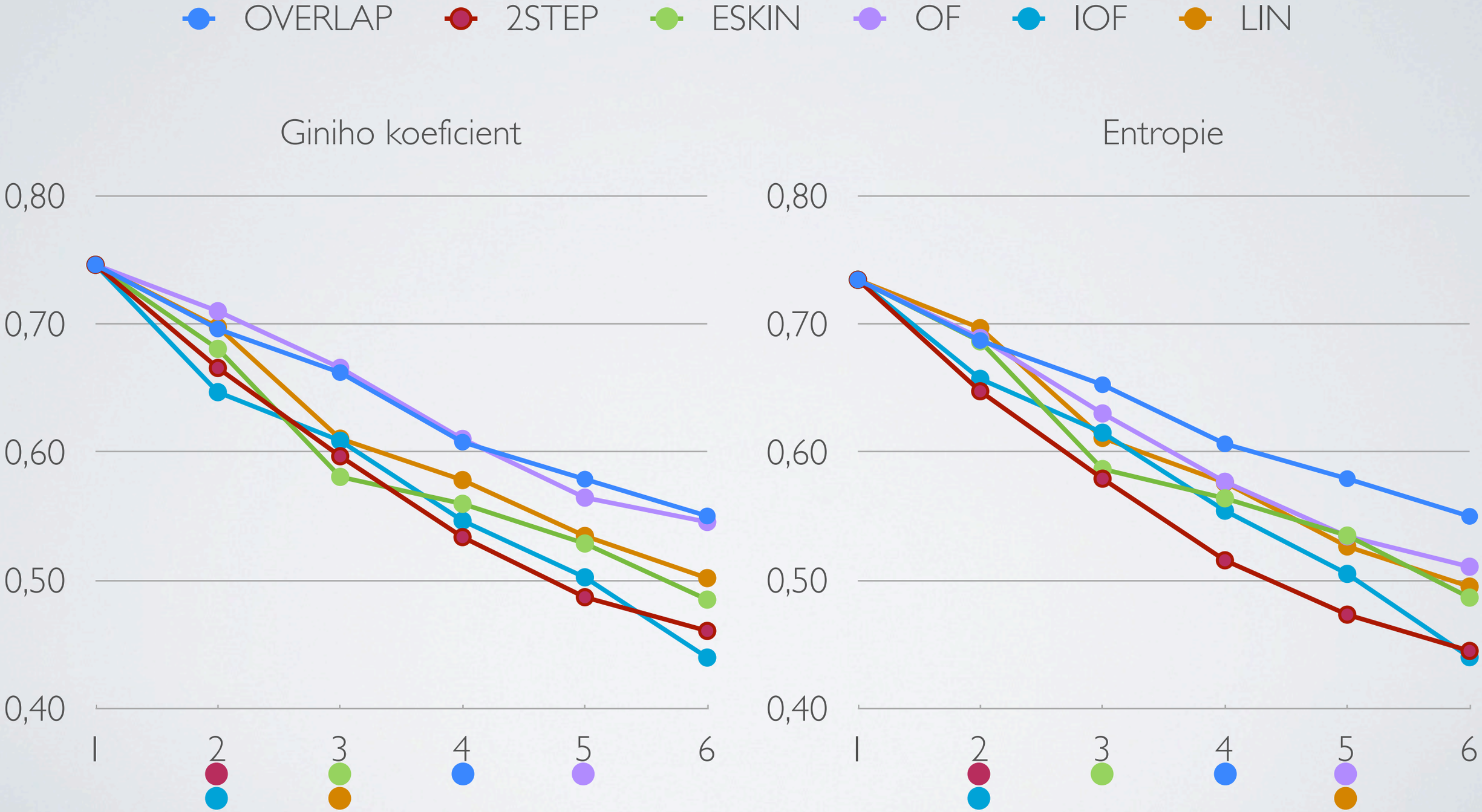
BREAST CANCER

OVERLAP 2STEP ESKIN OF IOF LIN





POST OPERATIVE



HODNOCENÍ SHLUKŮ

	Car	Hayes Roth	Breast Cancer	Post Operative	Průměr
overlap	3,5	5	4	5,5	4,5
2STEP	2	2	1	1,5	1,7
Eskin	1	3	6	3	3,3
OF	5,5	6	5	5,5	5,5
IOF	5,5	1	3	1,5	2,0
Lin	3,5	4	2	4	3,3

ZÁVĚR

- overlap dosahovala podprůměrných výsledků ve všech souborech
- IOF má výborné výsledky především u menších souborů
- Lin dosahuje dobrých výsledků u souborů s větším počtem kategorií
- Dvoukroková shluková analýza produkovala velmi dobré shluky ve všech souborech

DĚKUJI ZA POZORNOST