Schéma identifikační procedury

systém
S

generátor rekonstrukčních hypotéz

dekompozice na podsystémy
$=$ rekonstrukční hypotéza
$G = \{{}^1S, {}^2S, \ldots, {}^qS\}$

nestranné spojení

rekonstrukce
S^*

kvalita dekompozice

$S \overset{?}{\sim} S^* : \cancel{\Delta}(S, S^*) = \Delta G$

- G a S nejsou porovnatelné
  nelze srovnat kvalitu G a S
- S^* a S jsou porovnatelné

${}^2p(B, C)$

${}^1S \quad {}^2S$

${}^1p(A, B) \qquad A \cup B \qquad {}^1p(A \cup B)$

${}^1p(B) = {}^2p(B)$ lok.

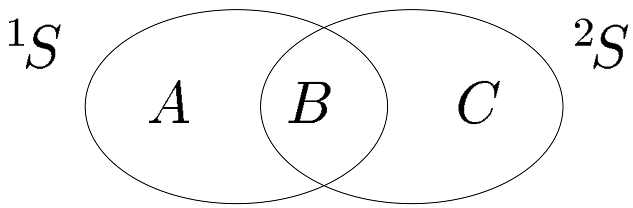
$S^* \qquad p^*(A, B, C)$

$p^*(A, B) = {}^1p(A, B)$
$p^*(B, C) = {}^2p(B, C)$

Spojení dvou marginálních funkcí přípustnosti

Marginální funkce přípustnosti:

rozklad množiny proměnných:



$$^1p_B \colon \mathcal{R}(A) \times \mathcal{R}(B) \to \langle 0, 1 \rangle$$
$$^2p_B \colon \mathcal{R}(B) \times \mathcal{R}(C) \to \langle 0, 1 \rangle$$

Spojení: $\quad {}^1p_B * {}^2p_B \colon \mathcal{R}(A) \times \mathcal{R}(B) \times \mathcal{R}(C) \to \langle 0, 1 \rangle$

Nestranné spojení (o maximální entropii):

$$p_B^*(A, B, C) = ({}^1p_B * {}^2p_B)(A, B, C) \stackrel{\text{def}}{=} {}^1p_B(A, B) \cdot {}^2p_B(C \mid B)$$

$^1p * {}^2p = {}^2p * {}^1p$

Pozn: ${}^1p_B(B) = {}^2p_B(B)$ (kompatibilita), ${}^1p_B * {}^2p_B = {}^2p_B * {}^1p_B$

Speciální případy:

$A = \emptyset: \quad ({}^1p_B * {}^2p_B)(B, C) = {}^1p_B(B) \cdot {}^2p_B(C \mid B) = {}^2p_B(B, C)$ nic se nestalo

$B = \emptyset: \quad ({}^1p_B * {}^2p_B)(A, C) = {}^1p_B(A) \cdot {}^2p_B(C)$ sdružené rozdělení nezávislých proměnných

Příklad

a — $\boxed{1S}$ — b — $\boxed{2S}$ — c

$A = \{a\}$, $B = \{b\}$, $C = \{c\}$

podsystémy:

1S

a	b	${}^1p_B(a,b)$
0	0	0.4
0	1	0.3
1	0	0.2
1	1	0.1

2S

b	c	${}^2p_B(b,c)$
0	0	0.4
0	1	0.2
1	0	0.1
1	1	0.3

lokální kompatibilita:

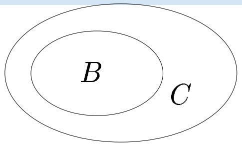
b	${}^1p_B(b)$	${}^2p_B(b)$
0	0.6	0.6
1	0.4	0.4

${}^2p(b,c)\,/\,{}^2p(b)$

- vyplníme tabulku přípustnosti stavu pro $p^*_B(a,b,c) = {}^1p_B(a,b) * {}^2p_B(c \mid b)$

a	b	c	$p^*_B(a,b,c)$
0	0	0	$0.4 \cdot 0.4/0.6 = 0.266$
0	0	1	$0.4 \cdot 0.2/0.6 = 0.133$
0	1	0	0.075
0	1	1	0.225
1	0	0	$\vdots$
1	0	1	
1	1	0	
1	1	1	

Poznámky



1. Připojením podsystému ‚se nestane nic'

$$({}^1p * {}^2p)(B,C) = {}^1p(B) \cdot {}^2p(C \mid B) = {}^2p(B) \cdot {}^2p(C \mid B) = {}^2p(B,C)$$

2. Je-li $p^*(A,B,C) = p(A,B) \cdot p(C \mid B)$, potom marginální funkce $p(A,B)$ a $p(B,C)$ jsou zachovány:

$$p^*(A,B) = \sum_{C_k} p^*(A,B,C_k) = \sum_{C_k} p(A,B) \cdot p(C_k \mid B) = p(A,B)$$

$$p^*(B,C) = \sum_{A_i} p^*(A_i,B,C) = \sum_{A_i} p(A_i,B) \cdot p(C \mid B) = p(B,C)$$

$p(A,B,C) \rightarrow p(A,B), \; p(B,C) \quad * \rightarrow p^*$

3. Marginální funkce $p(A,C)$ obecně zachována není.

$p^*(A,C) \overset{?}{=} p(A,C)$

a	b	c	p	p^*
0	0	0	0.0370	0.0106
0	0	1	0.1111	0.1376
0	1	0	0.0741	0.0684
0	1	1	0.3704	0.3761
1	0	0	0.0000	0.0265
1	0	1	0.3704	0.3439
1	1	0	0.0000	0.0057
1	1	1	0.0370	0.0313

a	c	p^*	p
0	0	0.0790	0.1111
0	1	0.5136	0.4815
1	0	0.0322	0.0000
1	1	0.3753	0.4074

pozn.

- zde $A = \{a\}$, $B = \{b\}$, $C = \{c\}$
- $p^*(a,b,c)$ – rekonstruována z $p(a,b)$, $p(b,c)$, získaných marginalizací z $p(a,b,c)$

důkaz tvrzení na následující straně (39)

$$-\sum_{i,j,k} \bar{p}(A_i,B_j,C_k)\log p(A_i,B_j)\cdot p(C_k|B_j)$$

$$-\sum_{i,j,k} \bar{p}(A_i,B_j,C_k)\log p(A_i,B_j) - \sum_{i,j,k} \bar{p}(A_i,B_j,C_k)\log p(C_k|B_j)$$

$$-\sum_{i,j} p(A_i,B_j)\cdot\log p(A_i,B_j) - \sum_{j,k} p(B_j,C_k)\log p(C_k|B_j)$$

$$1=\sum_{k} p(C_k|B_j)$$

$$p(C_k|B_j)\cdot p(B_j) \to \sum_{i} p(A_i,B_j)$$

!

$$-\sum_{i,j,k}\underbrace{p(A_i,B_j)\cdot p(C_k|B_j)}_{p^*(A_i,B_j,C_k)}\log p(A_i,B_j) - \sum_{i,j,k}\underbrace{p(C_k|B_j)\cdot p(A_i,B_j)}_{p^*}\log p(C_k|B_j)$$

$$= H(S^*)$$ QED

Věta o maximalitě entropie spojení

Nechť $\bar{S}$ s funkcí přípustnosti stavu $\bar{p}(A, B, C)$ je libovolná rekonstrukce systému S s funkcí přípustnosti stavu $p(A, B, C)$ a nechť S^* je rekonstrukce spojením $p^*(A, B, C) = p(A, B)\, p(C \mid B)$. Potom $H(\bar{S}) \leq H(S^*)$.

$p^*(A, B, C)$ $\quad \bar{p}(A, B, C) \rightarrow H(\bar{S})$

Pozn: Platí i pro případ, kdy $\bar{S}$ je původní systém S.

Důkaz

Předpokládejme, že existuje $\bar{p}$ tak, že $H(\bar{S}) > H(S^*)$. Použijeme Gibbsovu nerovnost pro rozdělení pravděpodobností

$$H(\bar{S}) = - \sum_{i,j,k} \bar{p}(A_i, B_j, C_k) \log \bar{p}(A_i, B_j, C_k) \leq - \sum_{i,j,k} \bar{p}(A_i, B_j, C_k) \log p^*(A_i, B_j, C_k) = H(S^*)$$

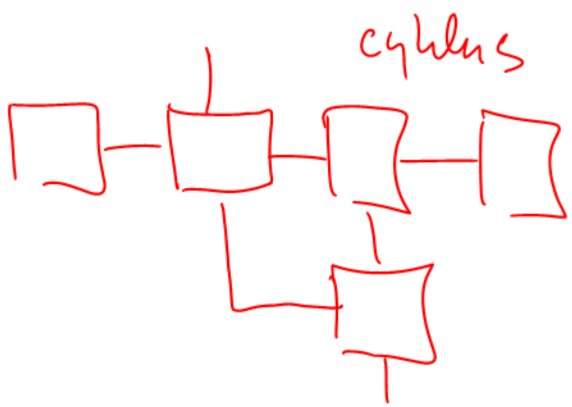
a to, že $p^*(A, B, C) = p(A, B) \cdot p(C \mid B)$, $\bar{p}(A, B) = p(A, B)$ a $\bar{p}(B, C) = p(B, C)$. Dostaneme $H(\bar{S}) \leq H(S^*)$, což je spor.

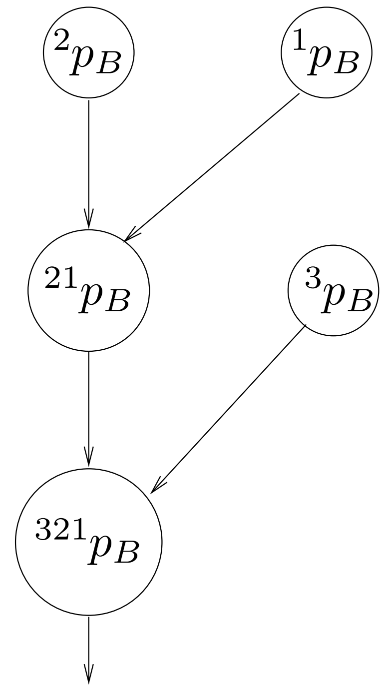
Spojovací procedura

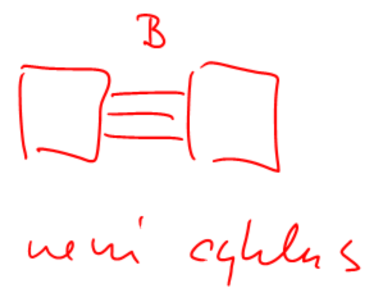
Jednoduchá procedura pro $\{{}^1S, {}^2S, \ldots, {}^qS\}$:

$$p_B^* = {}^q p_B * \left({}^{q-1} p_B * (\cdots * ({}^2 p_B * {}^1 p_B))\right)$$

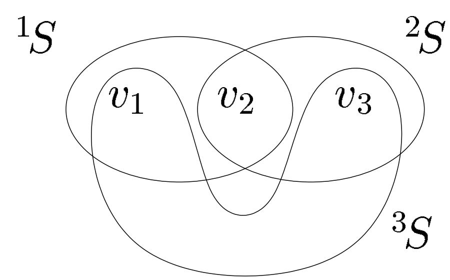
libovolné pořadí spojování







Problém: cyklické vazby mezi proměnnými

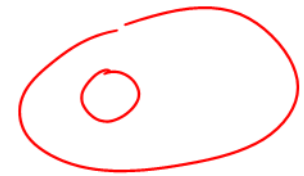


Je-li splněna podmínka $[\, p_B^* \downarrow {}^iS \,] = {}^i p_B$, potom je spojení nestranné (o maximální entropii). Jinak nutno použít iterativní proceduru (viz dále).

Iterativní spojovací procedura

Odstraní vliv cyklických vazeb.

Iterativní spojovací procedura $S^*({}^1S, {}^2S, \ldots, {}^qS; \varepsilon)$



1. $k := 0, \qquad p_0^* := {}^q p_B * {}^{q-1}p_B * \cdots * {}^1 p_B$
2. $k := k+1,\ p_k^* := {}^q p_B * {}^{q-1}p_B * \cdots * \left({}^1 p_B * p_{k-1}^*\right)$
3. jestliže $\max\limits_{s \in \mathcal{R}(S)} |p_k^*(s) - p_{k-1}^*(s)| < \varepsilon$ stop, jinak opakuj krok 2.

Jestliže $\sum\limits_{s \in \mathcal{R}(S)} p_k^*(s) = 1$, pak řešení nalezeno s chybou $\pm\varepsilon$, jinak $\{{}^1S, {}^2S, \ldots, {}^qS\}$ není globálně konzistentní.

Výsledkem je nestranná rekonstrukce s chybou ε

- nutná jen, pokud pro nějaké $i = 1, 2, \ldots, q$ platí $\left[\, p_B \downarrow {}^iS \,\right] \neq {}^i p_B$
- připojování nekonzistentních podsystémů ‚přepisováním marginálních funkcí'

$$ {}^2p_B(B) \cdot {}^1p_B(C \mid B) = \frac{{}^2p_B(B)}{{}^1p_B(B)} \cdot {}^1p_B(B, C) $$

- korektnost: $0 \leq p_k^* \leq 1$ protože každé spojení je násobení pravděpodobností, marginální funkce zachovány, ale neplatí nutně podmínka součtu do jedničky

Důkaz konvergence: [Brown, Chase, Pittenger. *Linear Algebra and Its Applications* 190:1-38, 1993]

Příklad 1

Máme tři podsystémy systému
s funkcí přípustnosti stavu p_{abc}

a	b	p_{ab}	b	c	p_{bc}	a	c	p_{ac}
0	0	0.1481	0	0	0.0370	0	0	0.1111
0	1	0.4444	0	1	0.4815	0	1	0.4815
1	0	0.3704	1	0	0.0741	1	0	0.0000
1	1	0.0370	1	1	0.4074	1	1	0.4074

Marginální funkce přípustnosti stavu po jednoduchém spojení $p_{ac} * (p_{ab} * p_{bc})$ p_{ac} se připojil poslední

a	b	$\hat{p}_{ab}$	b	c	$\hat{p}_{bc}$	a	c	$\hat{p}_{ac}$
0	0	0.1438	0	0	0.0149	0	0	0.1111
0	1	0.4487	0	1	0.5023	0	1	0.4815
1	0	0.3734	1	0	0.0962	0	0	0.0000
1	1	0.0340	1	1	0.3866	1	1	0.4074

Srovnání výsledku jednoduchého a iterativního spojení

a	b	c	p_{abc}	$p_{ac} * (p_{ab} * p_{bc})$	p^*_{abc}
0	0	0	0.0370	0.0149	0.0370
0	0	1	0.1111	0.0962	0.1111
0	1	0	0.0741	0.1290	0.0741
0	1	1	0.3704	0.3525	0.3704
1	0	1	0.3704	0.3734	0.3704
1	1	1	0.0370	0.0340	0.0370

Příklad 2

Máme tři podsystémy, které nejsou globálně konzistentní (viz str. 34)

v_1	v_2	${}^1p_B$	v_2	v_3	${}^2p_B$	v_1	v_3	${}^3p_B$
0	0	0.0000	0	0	0.0000	0	0	0.4000
0	1	0.7000	0	1	0.3000	0	1	0.3000
1	0	0.3000	1	0	0.7000	1	0	0.3000
1	1	0.0000	1	1	0.0000	1	1	0.0000

Výsledek jednoduchého i iterativního spojení

v_1	v_3	v_2	p_k
0	0	0	0.0000
0	0	1	0.4000
0	1	0	0.0000
0	1	1	0.0000
1	0	0	0.0000
1	0	1	0.0000
1	1	0	0.0000
1	1	1	0.0000

- podmínka $\sum_{s \in \mathcal{S}} p_k(s) = 1$ není splněna

Rekonstrukční chyba

R(S) R(S) $s = (a_i, b_j, c_k)$ v předchozí notaci

Dvě rozdělení $p^*(s)$, $p(s)$, $s \in \cancel{\mathcal{S}}$ nad stavovým prostorem $\cancel{\mathcal{S}}$.

Kullbackova-Leiblerova vzdálenost

(informační vzdálenost, poměrná entropie)

$$L(p, p^*) = \sum_{s \in \cancel{\mathcal{S}}} p(s) \log \frac{p(s)}{p^*(s)}$$

R(S)

Vlastnosti:

- $L(p, p^*) \geq 0$ (⇐ Gibbsova nerovnost)
- $L(p, p^*) \neq L(p^*, p)$ (není metrika)
- $L(p, p^*) = 0 \quad \Leftrightarrow \quad p \equiv p^*$
- je-li p^* nestranné spojení, pak

  1. $p^*(s) = 0 \;\Rightarrow\; p(s) = 0$
  2. (*1) $L(p, p^*) = H(S^*) - H(S)$
  3. ale $L(p^*, p) \neq H(S) - H(S^*)$

  $\cancel{H(S^*|S)}$

$p(A,B) \cdot p(C|B)$

1. $0 = p(A,B) = \sum_{c_k} p(A,B,c_k) \Rightarrow p(\underbrace{A,B,C}_{s}) = 0$
2. $0 = p(C|B) = \frac{1}{p(B)} \cdot p(B,C) \Rightarrow p(s) = 0$

Pro generativní systém

$$L(p, p^*) = \sum_{\overline{g} \in \overline{G}} p(\overline{g}) \sum_{g \in G} p(g \mid \overline{g}) \log \frac{p(g \mid \overline{g})}{p^*(g \mid \overline{g})}$$

$$= \cdots = H(G^* \mid \overline{G}^*) - H(G \mid \overline{G})$$

Rekonstrukční chyba: $\boxed{\Delta G = \Delta(S, S^*) = L(p, p^*)}$

Příklad

a	b	$p_B(a,b)$	$p^*_B(a,b) = {}^1p_B(a)\,{}^2p_B(b)$
0	0	0	0.21
0	1	0.3	0.09
1	0	0.7	0.49
1	1	0	0.21

a	${}^1p_B(a)$
0	0.3
1	0.7

b	${}^2p_B(b)$
0	0.7
1	0.3

$$L(S,S^*) = 0.3\log\frac{0.3}{0.09} + 0.7\log\frac{0.7}{0.49} = 0.6109$$

$$H(S^*) - H(S) = 1.2217 - 0.6109 = 0.6109$$

$p(a,b)$

a \ b	0	1
0	0	0.3
1	0.7	0

$p(a)$

0	0.3
1	0.7

$p(b)$

0	1
0.7	0.3

$p^*(a,b) = p(a)\cdot p(b)$

p^*

0.21	0.09
0.49	0.21

Test statistické významnosti struktury systému

$$H_0 : p(a_i, b_j, c_k) = ({}^1\!p * {}^2\!p)(a_i, b_j, c_k)$$

skutečná četnost $n(a_i, b_j, c_k)$

teoretická četnost $n^*(a_i, b_j, c_k) = n \cdot ({}^1\!p * {}^2\!p)(a_i, b_j, c_k)$

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n(a_i, b_j, c_k) - n^*(a_i, b_j, c_k)\right)^2}{n^*(a_i, b_j, c_k)}$$

- $I \cdot J \cdot K - J(I + K) + 1$ stupňů volnosti
- Lze jen u identifikace struktury systému, ne u prosté rekonstrukce

potřebujeme n i n^*, což máme jen při identifikaci struktury

poznámky

Počet stupňů volnosti

1. celý stavový prostor má velikost $I \cdot J \cdot K$
2. 1 podmínka
$$\sum_{i,j,k} p(a_i, b_j, c_k) = 1$$
3. $p(a, b)$ musí být marginální funkce: $I \cdot J - 1$ nezávislých vedlejších podmínek
   protože $\sum_{i,j} p(a_i, b_j) = 1$
4. $p(b, c)$ musí být marginální funkce: $J \cdot K - 1$ nezávislých vedlejších podmínek

$$I \cdot J \cdot K - 1 - (I \cdot J - 1) - (J \cdot K - 1) = I \cdot J \cdot K - J(I + K) + 1$$

Mood, AM – Graybill, FA – Boes, DC. *Introduction to the Theory of Statistics.* McGraw-Hill 1974, str. 454ff.