

Tato prezentace je součástí wiki-prezentace [Metoda GUHA a systém LISp-Miner](#)

Je dostupná z [této adresy](#)

Verze 11. 9. 2019

Asociační pravidla v datech s neúplnou informací - ignorování neúplné informace

Jan Rauch

Katedra informačního a znalostního inženýrství

Vysoká škola ekonomická

Asociační pravidla v datech s neúplnou informací - ignorování neúplné informace

- Matice dat s neúplnou informací
- Doplnění matice dat s neúplnou informací
- Devítipolní a čtyřpolní tabulka
- Ignorování neúplné informace
- Poznámky

Matice dat s neúplnou informací

Výsledek předzpracování – matice dat \mathcal{M}^X s neúplnou informací :

\mathcal{M}^X	A_1	A_2	A_3	\dots	A_P
o_1	1	X	5	\dots	1
o_2	X	1	X	\dots	3
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
o_{n-1}	2	X	2	\dots	2
o_n	X	3	1	\dots	X

X – kód chybějící hodnoty

Doplnění matice dat s neúplnou informací

\mathcal{M}^X	A_1	A_2	A_3	\dots	A_P
o_1	1	X	5	\dots	1
o_2	X	1	X	\dots	3
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
o_{n-1}	2	X	2	\dots	2
o_n	X	3	1	\dots	X

Matice dat \mathcal{M}^X

\mathcal{M}	A_1	A_2	A_3	\dots	A_P
o_1	1	3	5	\dots	1
o_2	3	1	4	\dots	3
\vdots	\vdots	\vdots	\vdots	\ddots	\vdots
o_{n-1}	2	2	2	\dots	2
o_n	2	3	1	\dots	1

Matice dat \mathcal{M}

Obrázek 4.6. Matice dat \mathcal{M}_X s neúplnou informací a příklad jejího doplnění \mathcal{M}

Každá chybějící hodnota se nahradí některou z přípustných hodnot

Devítipolní a čtyřpolní tabulka (1)

\mathcal{M}^X	φ	ψ
o_1	1	1
o_2	X	1
o_3	0	X
o_4	0	0
...
o_{n-1}	X	0
o_n	1	X



\mathcal{M}^X	ψ	ψ_X	$\neg\psi$
φ	$f_{1,1}$	$f_{1,X}$	$f_{1,0}$
φ_X	$f_{X,1}$	$f_{X,X}$	$f_{X,0}$
$\neg\varphi$	$f_{0,1}$	$f_{0,X}$	$f_{0,0}$

Devítipolní a čtyřpolní tabulka (2)

\mathcal{M}^X	φ	ψ
o_1	1	1
o_2	X	1
o_3	0	X
o_4	0	0
...
o_{n-1}	X	0
o_n	1	X



\mathcal{M}^X	ψ	ψ_X	$\neg\psi$
φ	$f_{1,1}$	$f_{1,X}$	$f_{1,0}$
φ_X	$f_{X,1}$	$f_{X,X}$	$f_{X,0}$
$\neg\varphi$	$f_{0,1}$	$f_{0,X}$	$f_{0,0}$



\mathcal{M}	ψ	$\neg\psi$
φ	$f_{1,1} + f_{1,X,a} +$ $+ f_{X,1,a} + f_{X,X,a}$	$f_{1,0} + f_{1,X,b} +$ $+ f_{X,0,b} + f_{X,X,b}$
$\neg\varphi$	$f_{0,1} + f_{X,1,c} +$ $+ f_{0,X,c} + f_{X,X,c}$	$f_{0,0} + f_{X,0,b} +$ $+ f_{0,X,d} + f_{X,X,d}$

Ignorování neúplné informace (1)

- Při zpracování se neúplná informace ignoruje v tom smyslu, že kód chybějící hodnoty X se považuje za speciální kategorii. Tato kategorie zůstává skryta a nelze ji použít při zadání relevantních koeficientů.
- Tento přístup vychází z přístupu uvedeného v článku [Introduction to arules – A computational environment for mining association rules and frequent item sets](#), viz též [arules: Mining Association Rules and Frequent Itemsets](#). V článku se uvádí:
- *„By default it is assumed that missing values do not carry information and thus all of the corresponding dummy items are set to zero. If the fact that the value of a specific attribute is missing provides information (e.g., a respondent in an interview refuses to answer a specific question), the domain expert can create for the attribute a category for missing values which then will be included in the transactions as its own dummy item.“*

Ignorování neúplné informace (2)

- Pokud tedy má atribut A pro některý objekt hodnotu X , tak každý základní booleovský atribut $A(\alpha)$ má při ignorování neúplné informace pro tento objekt hodnotu 0 .
- Tedy také, pokud má v případě pesimistického doplnění základní booleovský atribut $A(\alpha)$ hodnotu X , pak v případě ignorování neúplné informace má $A(\alpha)$ hodnotu 0 .
- Pokud má v případě pesimistického doplnění základní booleovský atribut $A(\alpha)$ hodnotu 1 nebo 0 , pak v případě ignorování neúplné informace má $A(\alpha)$ tutéž hodnotu.

Příklad: atribut Město má kategorie Praha, Brno, Ostrava, kód neúplné informace je X .

M	Město	Ignorování neúplné informace			Zabezpečené doplnění		
		Město(Praha)	Město(Brno)	Město(Praha, Brno)	Město(Praha)	Město(Brno)	Město(Praha, Brno)
o_1	Praha	1	0	1	1	0	1
o_2	Brno	0	1	1	0	1	1
o_3	Ostrava	0	0	0	0	0	0
o_4	X	0	0	0	X	X	X

Ignorování neúplné informace (3)

M	Zabezpečené doplnění				Ignorování neúplné informace			
	A(α)	B(β)	A(α) \wedge B(β)	A(α) \vee B(β)	A(α)	B(β)	A(α) \wedge B(β)	A(α) \vee B(β)
o_1	1	1	1	1	1	1	1	1
o_2	1	X	X	1	1	0	0	1
o_3	1	0	0	1	1	0	0	1
o_4	X	1	X	1	0	1	0	1
o_5	X	X	X	X	0	0	0	0
o_6	X	0	0	X	0	0	0	0
o_7	0	1	0	1	0	1	0	1
o_8	0	X	0	X	0	0	0	0
o_9	0	0	0	0	0	0	0	0

- Pokud je hodnota konjunkce/disjunkce základních booleovských atributů v zabezpečeném doplnění 1, pak je hodnota téže konjunkce/disjunkce při ignorování neúplné informace je také 1.
- Pokud je hodnota konjunkce/disjunkce základních booleovských atributů v zabezpečeném doplnění X nebo 0, pak je hodnota téže konjunkce/disjunkce při ignorování neúplné informace 0.
- Analogické tvrzení platí i pro dílčí cedenty a cedenty

Ignorování neúplné informace (4)

Devítipolní tabulka

\mathcal{M}^X	ψ	ψ_X	$\neg\psi$
φ	$f_{1,1}$	$f_{1,X} \rightarrow f_{1,0}$	
φ_X	$f_{X,1}$	$f_{X,X} \rightarrow f_{X,0}$	
$\neg\varphi$	$f_{0,1}$	$f_{0,X} \rightarrow f_{0,0}$	

Čtyřpolní tabulka při ignorování neúplné informace

\mathcal{M}	ψ	$\neg\psi$
φ	$f_{1,1}$	$f_{1,X} + f_{1,0}$
$\neg\varphi$	$f_{X,1} + f_{0,1}$	$f_{X,X} + f_{X,0} + f_{0,X} + f_{0,0}$

Poznámky (1)

Ignorování neúplné informace

\mathcal{M}^X	ψ	ψ_X	$\neg\psi$
φ	$f_{1,1}$	$f_{1,X} \rightarrow f_{1,0}$	
φ_X	$f_{X,1}$	$f_{X,X} \rightarrow f_{X,0}$	
$\neg\varphi$	$f_{0,1}$	$f_{0,X} \rightarrow f_{0,0}$	

Zabezpečené doplnění devítipolní tabulky

\mathcal{M}^X	ψ	ψ_X	$\neg\psi$
φ	$f_{1,1}$	$f_{1,X} \rightarrow f_{1,0}$	
φ_X	$f_{X,1}$	$f_{X,X} \rightarrow f_{X,0}$	
$\neg\varphi$	$f_{0,1}$	$f_{0,X} \rightarrow f_{0,0}$	

Existuje devítipolní tabulka, pro kterou platí:

- při ignorování neúplné informace je pravidlo $\varphi \Rightarrow_{0.9,30} \psi$ pravdivé
- při zabezpečeném doplnění je pravidlo $\varphi \Rightarrow_{0.9,30} \psi$ nepravdivé, viz následující slide.

Poznámky (2)

\mathcal{M}^X	ψ	ψ_X	$\neg\psi$
φ	90	5	5
φ_X	0	40	50
$\neg\varphi$	0	0	0

Ignorování neúplné informace

\mathcal{M}^X	ψ	$\neg\psi$
φ	90	10
$\neg\varphi$	0	90

$$\varphi \Rightarrow_{0.9,30} \psi$$

Zabezpečené doplnění devítipolní tabulky

\mathcal{M}^X	ψ	$\neg\psi$
φ	90	190
$\neg\varphi$	0	0

~~$$\varphi \Rightarrow_{0.9,30} \psi$$~~

Poznámky (3)

Pokud připouštíme, že se v matici dat může vyskytovat neúplná informace, ale v konkrétní matici dat \mathcal{A} není neúplná informace, pak devítipolní tabulka pro \mathcal{A} vypadá takto:

\mathcal{A}	ψ	ψ_X	$\neg\psi$
φ	u	0	v
φ_X	0	0	0
$\neg\varphi$	w	0	z

Ignorování neúplné informace

Zabezpečené doplnění

Optimistické doplnění

Vynechání

<table border="1"> <thead> <tr> <th>\mathcal{A}</th> <th>ψ</th> <th>$\neg\psi$</th> </tr> </thead> <tbody> <tr> <td>φ</td> <td>u</td> <td>v</td> </tr> <tr> <td>$\neg\varphi$</td> <td>w</td> <td>z</td> </tr> </tbody> </table>	\mathcal{A}	ψ	$\neg\psi$	φ	u	v	$\neg\varphi$	w	z	=	<table border="1"> <thead> <tr> <th>\mathcal{A}</th> <th>ψ</th> <th>$\neg\psi$</th> </tr> </thead> <tbody> <tr> <td>φ</td> <td>u</td> <td>v</td> </tr> <tr> <td>$\neg\varphi$</td> <td>w</td> <td>z</td> </tr> </tbody> </table>	\mathcal{A}	ψ	$\neg\psi$	φ	u	v	$\neg\varphi$	w	z	=	<table border="1"> <thead> <tr> <th>\mathcal{A}</th> <th>ψ</th> <th>$\neg\psi$</th> </tr> </thead> <tbody> <tr> <td>φ</td> <td>u</td> <td>v</td> </tr> <tr> <td>$\neg\varphi$</td> <td>w</td> <td>z</td> </tr> </tbody> </table>	\mathcal{A}	ψ	$\neg\psi$	φ	u	v	$\neg\varphi$	w	z	=	<table border="1"> <thead> <tr> <th>\mathcal{A}</th> <th>ψ</th> <th>$\neg\psi$</th> </tr> </thead> <tbody> <tr> <td>φ</td> <td>u</td> <td>v</td> </tr> <tr> <td>$\neg\varphi$</td> <td>w</td> <td>z</td> </tr> </tbody> </table>	\mathcal{A}	ψ	$\neg\psi$	φ	u	v	$\neg\varphi$	w	z
\mathcal{A}	ψ	$\neg\psi$																																								
φ	u	v																																								
$\neg\varphi$	w	z																																								
\mathcal{A}	ψ	$\neg\psi$																																								
φ	u	v																																								
$\neg\varphi$	w	z																																								
\mathcal{A}	ψ	$\neg\psi$																																								
φ	u	v																																								
$\neg\varphi$	w	z																																								
\mathcal{A}	ψ	$\neg\psi$																																								
φ	u	v																																								
$\neg\varphi$	w	z																																								

Díky vnitřní reprezentaci matice dat je v tomto případě procedura 4ft-Miner (podstatně) rychlejší při volbě *Ignorování neúplné informace*, než při ostatních způsobech zpracování neúplné informace.