

Vysoká škola báňská – Technická univerzita Ostrava

Fakulta elektrotechniky a informatiky

Katedra informatiky

Asociační pravidla (metoda GUHA)

Ing. Michal Burda (michal.burda@vsb.cz)

Získávání znalostí z dat

Brno, 27. ledna 2004

Úvod – metoda GUHA (General Unary Hypothesis Automaton)

- „automat na obecné unární hypotézy“
- původní česká metoda (Hájek, Havránek, Chytil, 80. léta 20. stol.)
- systematické vytváření hypotéz na základě empirických dat
- multidimenzionální asociační pravidla
- asociační, implikační a korelační kvantifikátory
- umí pracovat s neúplnou informací

Příklad

Vstupní data:

Věk	Plat	...	Auto
18–25	10 tis.	...	Škoda 120
41–60	40 tis.	...	BMW
⋮	⋮	⋮	⋮

Získané pravidlo:

$$\neg \text{věk}(1-17) \wedge \text{plat}(40 \text{ tis.}) \Rightarrow_{0,1;0,8} \text{Auto}(\text{BMW})$$

Pojmy

$$\neg \text{věk}(1-17) \wedge \text{plat}(40 \text{ tis.}) \Rightarrow_{0,1;0,8} \text{Auto(BMW)}$$

predikát – symbolické jméno veličiny

formule – predikáty složené pomocí logických spojek ($\wedge, \vee, \neg$)

kvantifikátor – symbol určující druh a intenzitu souvislosti

antecedent – formule na levé straně kvantifikátoru (předpoklad)

sukcedent – formule na pravé straně kvantifikátoru (závěr)

formální sentence – pravidlo

pravdivá sentence (hypotéza) – sentence, jejíž pravdivost je potvrzena vyhodnocením kvantifikátoru v datech

Pojmy

$$\neg \text{věk}(1-17) \wedge \text{plat}(40 \text{ tis.}) \Rightarrow_{0,1;0,8} \text{Auto(BMW)}$$

predikát – symbolické jméno veličiny

formule – predikáty složené pomocí logických spojek ($\wedge, \vee, \neg$)

kvantifikátor – symbol určující druh a intenzitu souvislosti

antecedent – formule na levé straně kvantifikátoru (předpoklad)

sukcedent – formule na pravé straně kvantifikátoru (závěr)

formální sentence – pravidlo

pravdivá sentence (hypotéza) – sentence, jejíž pravdivost je potvrzena vyhodnocením kvantifikátoru v datech

Pojmy

$$\neg \text{věk}(1-17) \wedge \text{plat}(40 \text{ tis.}) \Rightarrow_{0,1;0,8} \text{Auto}(\text{BMW})$$

predikát – symbolické jméno veličiny

formule – predikáty složené pomocí logických spojek ($\wedge, \vee, \neg$)

kvantifikátor – symbol určující druh a intenzitu souvislosti

antecedent – formule na levé straně kvantifikátoru (předpoklad)

sukcedent – formule na pravé straně kvantifikátoru (závěr)

formální sentence – pravidlo

pravdivá sentence (hypotéza) – sentence, jejíž pravdivost je potvrzena vyhodnocením kvantifikátoru v datech

Pojmy

$$\neg \text{věk}(1-17) \wedge \text{plat}(40 \text{ tis.}) \Rightarrow_{0,1;0,8} \text{Auto(BMW)}$$

predikát – symbolické jméno veličiny

formule – predikáty složené pomocí logických spojek ($\wedge, \vee, \neg$)

kvantifikátor – symbol určující druh a intenzitu souvislosti

antecedent – formule na levé straně kvantifikátoru (předpoklad)

sukcedent – formule na pravé straně kvantifikátoru (závěr)

formální sentence – pravidlo

pravdivá sentence (hypotéza) – sentence, jejíž pravdivost je potvrzena vyhodnocením kvantifikátoru v datech

Pojmy

$$\neg \text{věk}(1-17) \wedge \text{plat}(40 \text{ tis.}) \Rightarrow_{0,1;0,8} \text{Auto(BMW)}$$

predikát – symbolické jméno veličiny

formule – predikáty složené pomocí logických spojek ($\wedge, \vee, \neg$)

kvantifikátor – symbol určující druh a intenzitu souvislosti

antecedent – formule na levé straně kvantifikátoru (předpoklad)

sukcedent – formule na pravé straně kvantifikátoru (závěr)

formální sentence – pravidlo

pravdivá sentence (hypotéza) – sentence, jejíž pravdivost je potvrzena vyhodnocením kvantifikátoru v datech

Pojmy

$$\neg \text{věk}(1-17) \wedge \text{plat}(40 \text{ tis.}) \Rightarrow_{0,1;0,8} \text{Auto(BMW)}$$

predikát – symbolické jméno veličiny

formule – predikáty složené pomocí logických spojek ($\wedge, \vee, \neg$)

kvantifikátor – symbol určující druh a intenzitu souvislosti

antecedent – formule na levé straně kvantifikátoru (předpoklad)

sukcedent – formule na pravé straně kvantifikátoru (závěr)

formální sentence – pravidlo

pravdivá sentence (hypotéza) – sentence, jejíž pravdivost je potvrzena vyhodnocením kvantifikátoru v datech

Pojmy

$$\neg \text{věk}(1-17) \wedge \text{plat}(40 \text{ tis.}) \Rightarrow_{0,1;0,8} \text{Auto}(\text{BMW})$$

predikát – symbolické jméno veličiny

formule – predikáty složené pomocí logických spojek ($\wedge, \vee, \neg$)

kvantifikátor – symbol určující druh a intenzitu souvislosti

antecedent – formule na levé straně kvantifikátoru (předpoklad)

sukcedent – formule na pravé straně kvantifikátoru (závěr)

formální sentence – pravidlo

pravdivá sentence (hypotéza) – sentence, jejíž pravdivost je potvrzena vyhodnocením kvantifikátoru v datech

Kvantifikátory (1)

Důležité je dávat na výstup pouze sentence, které jsou nějak zajímavé, které podporují nějakou **hypotézu**.

Kvantifikátory popisují druh a intenzitu takové hypotézy. Postihují tedy druh a sílu vazby mezi antecedentem a sukcedentem.

Druhy kvantifikátorů:

- implikační – A (asi, většinou) je příčinou B ($A \Rightarrow B$)
- asociační – A (asi, většinou) souvisí s B ($A \sim B$)
- korelační – za podmínky F spolu hodnoty A a B (asi, většinou) korelují ($A \text{ corr } B/F$)

Čtyřpolní tabulky

Síla vztahu určeného daným kvantifikátorem se nestanovuje z jednotlivých hodnot datové tabulky, ale z vhodných **charakteristik** dat.

GUHA jako charakteristiky používá **frekvence**.

Frekvenční (čtyřpolní) tabulka:

	antecedent	$\neg$ antecedent	
sukcedent	a	b	r
$\neg$ sukcedent	c	d	s
	k	l	m

(kde $r = a + b$, $s = c + d$, $k = a + c$, $l = b + d$ a $m = a + b + c + d$)

Příklad výpočtu čtyřpolní tabulky

tequila	citróny
0	0
1	1
0	1
1	1
0	0
1	1
0	1
0	1
1	1
0	0
1	0
0	0

Sentence:

 $\text{tequila} \Rightarrow \text{citróny}$

	tequila	$\neg \text{tequila}$	
citróny	4	3	7
$\neg \text{citróny}$	1	4	5
	5	7	12

Příklad výpočtu čtyřpolní tabulky

tequila	citróny
0	0
1	1
0	1
1	1
0	0
1	1
0	1
0	1
1	1
0	0
1	0
0	0

Sentence:

 $\text{tequila} \Rightarrow \text{citróny}$

	tequila	$\neg \text{tequila}$	
citróny	4	3	7
$\neg \text{citróny}$	1	4	5
	5	7	12

Příklad výpočtu čtyřpolní tabulky

tequila	citróny
0	0
1	1
0	1
1	1
0	0
1	1
0	1
0	1
1	1
0	0
1	0
0	0

Sentence:

tequila $\Rightarrow$ citróny

	tequila	$\neg$ tequila	
citróny	4	3	7
$\neg$ citróny	1	4	5
	5	7	12

Příklad výpočtu čtyřpolní tabulky

tequila	citróny
0	0
1	1
0	1
1	1
0	0
1	1
0	1
0	1
1	1
0	0
1	0
0	0

Sentence:

 $\text{tequila} \Rightarrow \text{citróny}$

	tequila	$\neg \text{tequila}$	
citróny	4	3	7
$\neg \text{citróny}$	1	4	5
	5	7	12

Kvantifikátory (2)

Každý kvantifikátor je definován jako funkce frekvencí a, b, c, d . Pokud je výsledkem zobrazení hodnota 1, zkoumaná sentence je přijata jako **pravdivá**.

Implikační kvantifikátory (1)

A (asi, většinou) je příčinou B ($A \Rightarrow B$)...

1. $\Rightarrow_{s,p}$ **fundovaná implikace** (pro $s \in (0; 1)$ a $p > 0$):

$$\Rightarrow_{s,p} (a, b, c, d) = 1, \quad \text{je-li} \quad a \geq p \quad \text{a} \quad a \geq s(a + b)$$

Představuje požadavek na dosažení minimální podpory (p) a spolehlivosti (s) v datech.

2. $\Rightarrow_{s,p,\alpha}^!$ **dolní kritická implikace** (pro $s \in (0; 1)$, $p > 0$ a $\alpha \in \langle 0; 0,5 \rangle$):

$$\Rightarrow_{s,p,\alpha}^! (a, b, c, d) = 1, \quad \text{je-li} \quad \sum_{i=a}^r \binom{r}{i} s^i (1-s)^{r-i} \leq \alpha$$

Je založena na statistickém testu nulové hypotézy, že podmíněná pravděpodobnost sukcedentu za podmínky antecedentu je menší nebo rovna s , proti alternativní hypotéze, že je větší než s . Jde o test na hladině významnosti α . Hodnota 1 indikuje přijetí alternativní hypotézy.

Implikační kvantifikátory (2)

3. $\Rightarrow_{s,p,\alpha}^?$ **horní kritická implikace** (pro $s \in (0; 1)$, $p > 0$ a $\alpha \in \langle 0; 0,5 \rangle$):

$$\Rightarrow_{s,p,\alpha}^? (a, b, c, d) = 1, \quad \text{je-li} \quad \sum_{i=0}^a \binom{r}{i} s^i (1-s)^{r-i} > \alpha$$

Je založena na testu nulové hypotézy, že podmíněná pravděpodobnost sukcedentu za podmínky antecedentu je větší nebo rovna s , proti alternativní hypotéze, že je menší než s . Jde o test na hladině významnosti α . Hodnota 1 indikuje nezamítnutí nulové hypotézy.

Dolní a horní kritická implikace se liší pouze v tom, kterou statistickou chybu omezují hodnotou α .

Nechť $R_M(\Rightarrow^*)$ je množina všech pravdivých sentencí s implikačním kvantifikátorem $\Rightarrow^*$ v datech M ; pak platí:

$$R_M(\Rightarrow_{s,p,\alpha}^!) \subseteq R_M(\Rightarrow_{s,p}) \subseteq R_M(\Rightarrow_{s,p,\alpha}^?)$$

Asociační kvantifikátory (1)

A (asi, většinou) souvisí s B ($A \sim B$)...

1. $\sim_\delta$ **prosté vychýlení** (pro lib. $\delta \geq 0$)

$$\sim_\delta (a, b, c, d) = 1, \quad \text{je-li} \quad ad > e^\delta bc$$

(speciálně pro $\delta = 0$ dostáváme $ad > bc$).

2. $\sim_\alpha^1$ **Fisherův kvantifikátor** (pro lib. $\alpha \in \langle 0; 0,5 \rangle$):

$$\sim_\alpha^1 (a, b, c, d) = 1, \quad \text{je-li} \quad ad > bc \quad \text{a} \quad \sum_{i=a}^{\min(r,k)} \frac{\binom{k}{i} \binom{m-k}{r-i}}{\binom{m}{r}} \leq \alpha$$

Je založen na statistickém testu hypotézy o nezávislosti veličin proti alternativě o jejich kladné závislosti na hladině významnosti α . Hodnota 1 indikuje přijetí alternativní hypotézy.

Asociační kvantifikátory (2)

3. $\sim_{\alpha}^2$ χ^2 -kvantifikátor (pro $\alpha \in (0; 0,5)$):

$$\sim_{\alpha}^2(a, b, c, d) = 1, \quad \text{je-li} \quad ad > bc \quad \text{a} \quad \frac{(ad - bc)^2}{r k l s} m \geq \chi_{\alpha}^2$$

kde χ_{α}^2 je $(1 - 2\alpha)$ -kvantil χ^2 -rozložení s jedním stupněm volnosti. Tento kvantifikátor má stejné statistické pozadí jako Fisherův.

(Heuristická) pravidla pro použití jednotlivých kvantifikátorů (lh je součet největších délek antecedentu a sukcedentu, m počet záznamů):

- Pro χ^2 -kvantifikátor by mělo platit: $5 \times 2^{lh} \leq m$
- Pro Fisherův kvantifikátor by mělo platit: $2^{lh} \leq m$
- χ^2 -test má větší sílu než Fisherova statistika. Proto platí-li nerovnost $\min\{5 \times 2^{lh}, 250\} \leq m$, používáme raději kvantifikátor χ^2 místo Fisherova.

Korelační kvantifikátory (1)

Za podmínky F hodnoty A a B (asi, většinou) korelují ($A \text{ corr } B / F$)...
Všechny korelační kvantifikátory původní metody GUHA jsou založeny na pojmu pořadí.

Předpokládejme, že máme data vzniklá pozorováním dvou reálněhodnotových veličin, t_1 a t_2 . $R(i)$ definujeme jako počet objektů, pro něž hodnota veličiny t_1 je menší než hodnota t_1 pro i -tý objekt. Podobně definujeme $Q(i)$ (vzhledem k veličině t_2).

1. $s\text{-corr}_\alpha$ **Spearmanův kvantifikátor** (pro $\alpha \in (0; 0,5)$):

$$s\text{-corr}_\alpha(\langle t_1, t_2 \rangle) = 1, \quad \text{je-li} \quad \sum_{i=1}^m R(i)Q(i) \geq k_\alpha$$

kde k_α je vhodná konstanta. Za jistých dodatečných předpokladů jde o statistický test nulové hypotézy nezávislosti proti alternativní hypotéze o kladné závislosti. Hodnota 1 indikuje kladnou závislost na hladině α .

Korelační kvantifikátory (2)

2. $k\text{-corr}_\alpha$ **Kendallův kvantifikátor** (pro $\alpha \in (0; 0,5)$):

$$k\text{-corr}_\alpha(\langle t_1, t_2 \rangle) = 1,$$

$$\text{je-li } \sum_{i \neq j} \left(\text{sign}(R(i) - R(j)) \text{sign}(Q(i) - Q(j)) \right) \geq k_\alpha$$

kde k_α je vhodná konstanta. Statistická interpretace je stejná jako výše. $\text{sign}(x)$ znamená funkci **signum** (kladného čísla je 1, záporného -1).

3. $e\text{-corr}$ **pořadově ekvivalenční kvantifikátor**:

$$e\text{-corr}(\langle t_1, t_2 \rangle) = 1, \quad \text{je-li } R(i) = Q(i) \quad \text{pro } i = 1, \dots, m$$

GUHA procedury

Původní metoda GUHA se skládá z řady procedur, z nichž některé patří spíše do metod předzpracování.

V dalším se zaměříme na:

- ASSOC – vyhledávání sentencí s asociačními kvantifikátory
- IMPL – vyhledávání implikací
- CORREL – vyhledávání vysokých podmíněných korelací

Metoda ASSOC (1)

Vyhledávání sentencí s asociačními kvantifikátory. (Hledání těch dvojic antecedent/sukcedent, u kterých shody převažují nad neshodami.)

$$p_1 \wedge p_2 \wedge p_3 \sim^* p_4 \wedge p_5 \wedge p_6$$

Vstupy (tvar zkoumaných sentencí):

- maximální povolená délka antecedentu a sukcedentu
- které predikáty a s jakým znamením se mohou vyskytovat v antecedentu a které v sukcedentu
- druh kvantifikátoru

Výstupy – sentence, které splňují:

- jsou pravdivé v datech a vyhovují tvaru zadanému na vstupu
- antecedent i sukcedent jsou elementární konjunkce

Metoda ASSOC (2)

Nástin algoritmu:

1. Začíná se od sentencí tvořených jedním predikátem v antecedentu a jedním v sukcedentu. Postupně se testuje pravdivost všech přípustných kombinací počítáním čtyřpolních tabulek a vyhodnocováním funkce kvantifikátoru.
2. Následuje prodlužování antecedentu i sukcedentu. Obě formule jsou tvořeny jako elementární konjunkce. Opět se testuje pravdivost všech přípustných kombinací.
3. Délka sentence se zvětšuje až do maximální stanovené velikosti.

Metoda IMPL

Hledání pravdivých sentencí ve tvaru implikace. (Vyhledávání vysokých podmíněných pravděpodobností ve dvouhodnotových datech.)

$$p_1 \wedge p_2 \wedge p_3 \Rightarrow^* p_4 \vee p_5 \vee p_6$$

Vstupy (tvar zkoumaných sentencí):

- stejné jako u metody ASSOC (povolené predikáty, kvantifikátor a max. délka antecedentu a sukcedentu)

Výstupy – sentence, které splňují:

- jsou pravdivé v datech a vyhovují tvaru zadanému na vstupu
- antecedent je tvaru elementární konjunkce a sukcedent je tvaru elementární disjunkce

Algoritmus – podobný jako u ASSOC

Dedukční pravidla

Při hledání pravdivých sentencí metodami ASSOC a IMPL se dá využít různých dedukčních pravidel a tím urychlit celý proces, protože nemusíme testovat sentence, které lze odvodit z ostatních:

1. Pravidlo záměny ekvivalentních formulí.
2. Pravidlo úprav elementární implikace.
3. Pravidlo symetrie.
4. Pravidlo konzervativního zlepšování.
5. Pravidlo ostrého zlepšování pro konjunktivní asociace.

Neúplná informace

- Místo s dvouhodnotovými daty se pracuje s trojhodnotovými (0, X a 1, kde X znamená „nevyplněno“).
- Vyhodnocování formulí a funkcí kvantifikátorů se potom provádí tak, že se vždy bere v úvahu nejhorší možný způsob doplnění skutečných hodnot za X. Máme tak jistotu, že získaná pravidla by skutečně platila i v datech bez neúplné informace.

Procedura CORREL (1)

Hledání elementárních konjunkcí, pro které je podmíněná korelace dvou vybraných reálných veličin v datech vysoká.

$$(p_1 \text{ corr } p_2) / f$$

Vstupy:

- reálněhodnotové veličiny p_1 a p_2 , které chceme zkoumat
- užitý korelační kvantifikátor
- které predikáty a s jakým znamením se mohou vyskytovat v podmínce
- maximální povolená délka podmínky

Procedura CORREL (2)

Výstupy – sentence, které splňují:

- jsou pravdivé v datech
- p_1 , p_2 a korelační kvantifikátor corr jsou neměnné
- podmínka f je elementární konjunkce a vyhovuje tvaru zadanému na vstupu

Algoritmus hledání pravdivých sentencí spočívá v postupném generování podmínek povoleného tvaru, v jejich prodlužování až do zadané maximální velikosti a testováním pravdivosti vzniklých sentencí.

Statistický pohled (1)

- GUHA často používá pro rozhodnutí o pravdivosti sentence statistické testy.
- Statistika pracuje s pravděpodobností – vždy $\exists$ určitá pravděpodobnost, že je vyvozený závěr neplatný.
- **Simultánní statistická inference** – pravděpodobnost chyby v rozhodnutí o pravdivosti každé sentence je $\leq \alpha_i$. Celková pravděpodobnost chyby v alespoň jedné sentenci $s \in M$ je tedy nutně $\leq \sum_M \alpha_i$.
 $\Rightarrow$ pravděpodobnost, že mezi získanými asociačními pravidly je alespoň jedno nepravdivé, se blíží k jistotě.

Statistický pohled (2)

Má tedy získávání asociačních pravidel vůbec smysl?

Samozřejmě ano. Na vytěžené znalosti ovšem nemůžeme pohlížet jako na „jisté pravdy“. Nelze je prohlásit za zákony zkoumané oblasti.

Asociační pravidla nám pouze dávají přehled o charakteru dat. Ukazují, které **hypotézy** jsou daty podporovány, jaké vztahy jsou **možná** pravdivé a které směry dalšího bádání v datech jsou nadějně.

Konec. Čas na Vaše dotazy...