6. ASOCIACE

Čas ke studiu: 6 hodin

Cíl Po prostudování této kapitoly budete

- vědět, co jsou asociace mezi podmnožinami atributů a jaké typy asociací rozeznáváme,
- umět pro každý typ asociací popsat a zdůvodnit jejich výpočet,
- pomocí popsaných metod vybrat vhodnou metodu, analyzovat a řešit praktické úlohy generování asociací různých typů.

Výklad

6.1. Asociace jako vztahy mezi atributy

Jak víme, zkoumané objekty jsou popsány svými atributy. Jednou velkou skupinou, nejčastěji využívanou při dolování znalostí, je hledání vztahů mezi některými podmnožinami atributů.

Představme si objekty v databázi jako „nosiče“ mnoha vlastností. Může se stát, že mezi vlastnostmi jsou jisté, ne vždy zřetelné vztahy. Pokud experta nebo analytika „napadne“, že by mezi atributy A a B mohl být vztah například typu „jestliže A=1, pak B=5“, může si tuto vlastní hypotézu v datech otestovat. Ovšem, pokud takový předpoklad neformuluje, také ho neotestuje. Přitom mohou existovat zajímavé vztahy, které nikdo nepozná, protože nikoho nenapadnou.

Prapůvodem metody hledání asociací je metoda GUHA - původní česká metoda pražské skupiny autorů (P. Hájek, M. K. Chytil, T. Havránek) z roku 1964. Od té doby se rozšířila po celém světě.

Metoda používá prostředků matematické statistiky a matematické logiky. Automaticky a systematicky generuje a ověřuje hypotézy těchto typů:

1. A (asi, většinou) souvisí s B
2. A (asi, většinou) je příčinou B
3. za podmínky P veličiny A, B (asi, většinou) korelují

Pojmy „souvisí“, „je příčinou“, „korelují“ budou upřesněny níže, veličinami A, B rozumíme podmnožiny atributů zkoumaných objektů (původní, filtrované, odvozené)

Později, zvláště se začátkem využívání v marketingu, se rozvinuly i varianty základních asociací. Princip automatického generování a testování všech možných hypotéz však zůstává stejný – to je podstata rozdílu proti testování v matematické statistice.

Asociacemi tedy nazýváme vztahy mezi podmnožinami atributů. Rozlišujeme asociace

- **klasické** - mezi dvěma podmnožinami atributů
- **transakční** - v rámci velmi rozsáhlé množiny atributů
- **agregované** - mezi podmnožinou atributů a jejich skupinovými charakteristikami

Shrnutí pojmů 6.1.

Asociace jako vztahy mezi podmnožinami atributů.

Typy asociací.

6.2. Asociace základní

Cíl Po prostudování této kapitoly budete

- vědět, co jsou základní asociace mezi podmnožinami atributů,
- umět popsat jejich význam, interpretaci,
- pro praktické úlohy navrhnout smysluplné typy úloh na asociace

Výklad

Základní pojmy

Mějme binární datovou matici **X** s objekty {O1, O2, …} v řádcích a atributy {A, B,…, X, Y, …} ve sloupcích.

X

	A	**B**	**…**	**X**	**Y**	**…**
O1	x11	x12		…	…	
	…	…				
Oi	xi1	xi2				
	…	…				
Om	xm1	xm2				

Jména základních i složených veličin Aj nazveme **formule**, označíme **Fi**.

Z binárních formulí je možno vytvářet **složené binární formule** pomocí logických spojek negace, konjunkce a disjunkce.

Jsou-li Fi, Fj dvouhodnotové formule, pak také

$$\neg Fi \qquad Fi \wedge Fj \qquad Fi \vee Fj$$

jsou dvouhodnotové formule.

Pro asociace mají význam **elementární konjunkce**, tj. formule tvaru

$$\pm F1 \wedge \pm F2 \wedge \ldots \wedge \pm Fk$$

a **elementární disjunkce** tvaru

$$\pm F1 \vee \pm F2 \vee \ldots \vee \pm Fk$$

kde symbol ± znamená, že před Fi je negace nebo nic; Fi jsou navzájem různé veličiny.

Příklad 6.1.

Pro zdrojovou datovou matici Pacient (jméno, věk, váha, tlak, …, kouří, pije_alkohol, užívá_drogy, vegetarián, …, vysoký_tlak, rakovina_plic, cukrovka, infarkt, …) *jsou sledovanými objekty pacienti s některou z evidovaných nemocí.*

Jednoduché formule jsou například: kouří *[A/N],* vegetarián *[A/N],*

složené formule jsou například: pije_alkohol ∧ ¬kouří ∨ vegetarián.

Elementární konjunkcí je například: cukrovka ∧ ¬infarkt ∧ vysoký_tlak

Elementární disjunkcí je například: kouří ∨ pije_alkohol ∨ užívá drogy

Vztahy, které mohou lékaře zajímat, jsou například vztahy mezi atributy charakterizujícími styl života na jedné straně a atributy popisujícími diagnózu na straně druhé.

♦

❑ Charakteristiky binárních dat

Pro binární a kategoriální data užitečnou informaci shrnují charakteristiky **frekvence**. Obvykle je zapisujeme do tvaru frekvenční tabulky.

Pro data **X** a z nich vytvořené binární formule F1 a F2 má frekvenční (čtyřpolní) tabulka tvar

platí F1 \ F2	ano	ne	
ano	a	b	r
ne	c	d	s
	k	l	m

Frekvence ještě nedávají dostatečnou informaci o datech, proto se zavádí další charakteristiky definované na matici dat pomocí frekvencí a dávající hodnoty z oboru reálných čísel. Nazýváme je zobecněné **kvantifikátory.** Ty jsou vypočteny z hodnot čtyřpolní tabulky. Různé kvantifikátory dávají pak charakterizují různé typy vztahů mezi formulemi F1, F2.

Pomocí kvantifikátoru q se ze dvou formulí vytvoří formální **sentence**

F1 q F2

Formuli na levé straně sentence nazýváme **antecedentem**, na pravé straně **sukcedentem.**

Sentence **je pravdivá** na datech **X**, jestliže po aplikaci zobrazení definujícího kvantifikátor na tabulku dostaneme hodnotu 1. Sentence pravdivé na datech nazveme **hypotézami.**

Množinu sentencí pravdivých v datech **X** (množinu hypotéz) označíme **H(X).**

Pro danou matici dat **X** je dále dána množina sentencí (popsaná ne výčtem, ale pravidly, jak je generovat), které jsou považovány z nějakých důvodů za **relevantní otázky**. Množinu relevantních otázek označíme **RO**.

Zajímá nás, které sentence z RO jsou pravdivé v datech **X**.

Než uvedeme příklad na sentence, zavedeme si používané kvantifikátory.

❑ Typy kvantifikátorů

V úvodu kapitoly o asociacích jsme si řekli, že budeme hledat vztahy typu A (asi) souvisí s B, A(asi) je příčinou B, případně za podmínky P veličiny A, B spolu korelují. Pro tyto tři typy vztahů nyní zavedeme následující tři typy kvantifikátorů:

I. Kvantifikátory asociační pro binární veličiny

II. Kvantifikátory implikační pro binární veličiny

III. Kvantifikátory korelační pro reálné veličiny

IV.

- **Kvantifikátory asociační pro binární veličiny**

odpovídají intuitivní představě souvislosti: kvantifikátor q je asociační, jestliže sentence **F1 q F2** říká, že F1 souvisí s F2 (že shody jejich hodnot převažují nad neshodami).

Definice 6.1.

Kvantifikátor q je asociační, jestliže pro každé a, b, c, d takové, že q(a,b,c,d) = 1 a každé a'≥ a, b'≤b , c' ≤ c, d' ≥ d je q (a',b',c',d') = 1.

Asociační kvantifikátory jsou symetrické, tj. q(a, b, c, d) = q(a, c, b, d).

Jednotlivé asociační kvantifikátory jsou definovány takto:

1. prosté vychýlení $\sim\delta$ pro libovolné $\delta > 0$

$\sim\delta$ **(a,b,c,d) = 1** je-li ad > e^{δ} . bc … ad / bc > e^{δ}

2. Fisherův kvantifikátor $\sim\alpha 1$ pro libovolné $\alpha \in (0, 0.5>$

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

Tento kvantifikátor je založen na testu hypotézy o nezávislosti testovaných veličin proti alternativě o jejich kladné závislosti; jde o test na hladině významnosti α. Hodnota 1 indikuje přijetí alternativní hypotézy.

Je vhodný pro malá **m ∈ <20, 40>** nebo pro četnosti **a, b, c, d < 5**.

3. χ^2 **- kvantifikátor** $\sim\alpha^2$ pro libovolné $\alpha \in (0, 0.5>$

$\sim\alpha$ **(a,b,c,d) = 1** je-li ad > bc a $\frac{(ad-bc)}{rkls} m \ge \chi^2_\alpha$

kde χ_α^2 je (1 - α) - kvantil χ^2 rozložení s jedním stupněm volnosti; je to test asymptotický. Statistický význam má stejný, jako Fisherův test.

Je vhodný pro **m > 40** nebo pro **m ∈ <20, 40>** a četnosti **a, b, c, d > 5**.

Charakteristiky pro souvislosti (též asociace v užším slova smyslu jako "skoroekvivalence") odpovídají intuitivní představě souvislosti: kvantifikátor q je asociační, jestliže sentence říká, že antecedent souvisí se sukcedentem, že počet shod převládá nad počtem neshod.

Pokud ve čtyřpolní tabulce platí c=b=0, pak jde o logickou ekvivalenci.

• Kvantifikátory implikační pro binární veličiny

Je-li q kvantifikátor implikační, pak formule **F1 q F2** říká, že skoro všechny objekty splňující F1 splňují i F2.

Definice 6.2.

Kvantifikátor q je implikační, jestliže pro každé a, b, c, d takové, že q(a,b,c,d) = 1 a pro každé a' ≥ a, b' ≤ b a libovolné c', d' platí

$$q(a', b', c', d') := 1.$$

Implikační kvantifikátor není symetrický. Kvantifikátory implikační jsou definovány takto:

1. Fundovaná implikace $\Rightarrow_{P,S}$ (pro $S \in (0,1]$ a $P > 0$)

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

kde S ... minimální spolehlivost vztahu definována vhodnými charakteristikami
P ... minimální podpora, počet případů, pro něž je nalezená hypotéza platná

Pro S = 1 a P = 0 dostáváme obvyklou logickou implikaci

2. Dolní kritická implikace $\Rightarrow_{P,S,\alpha}$ (pro P, S jako výše, $\alpha \in (0, 0.5]$

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

je založen na testu nulové hypotézy, že podmíněná pravděpodobnost sukcedentu za podmínky antecedentu je menší nebo rovna P proti alternativní hypotéze, že je větší než S; jde o test na hladině významnosti α; hodnota 1 indikuje přijetí alternativní hypotézy. Usuzujeme tedy, že **přítomnost antecedentu zvyšuje pravděpodobnost platnosti sukcedentu.**

3. Horní kritická implikace $\Rightarrow^{P,S,\alpha}$ (pro P, S, α jako výše)

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

je založen na testu nulové hypotézy, že podmíněná pravděpodobnost sukcedentu za podmínky antecedentu je větší nebo rovna p proti alternativní hypotéze, že je menší než s; jde o test na hladině významnosti α; hodnota 1 indikuje nezamítnutí nulové hypotézy. Usuzujeme tedy, že **nelze vyloučit, že přítomnost antecedentu zvyšuje pravděpodobnost platnosti sukcedentu**.

• Korelační kvantifikátory pro reálná data

posuzují kladné závislosti dvou reálných veličin. Asociace používají korelační kvantifikátory založené na pojmu pořadí.

Předpokládejme, že data **X** obsahují objekty O1, O2, . . ., Om a pro ně dva reálné atributy t1 a t2. Dále, že pro žádné dva objekty O1, O2 neplatí O1(t1) = O2(t2) nebo O2(t2) = O1(t1). Pro libovolný objekt Oi definujme **pořadí R(i)** objektu vzhledem k t1 jako počet prvků množiny {Oj Î **X** | Oj(t1) £ Oi(t2)}. Podobně definujeme pořadí objektů Q(i) vzhledem k t2. Pak

1. **Spearmanův kvantifikátor** s-corr$_\alpha$ pro $\alpha \in (0, 0.5>$

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

kde k_α je vhodná konstanta. Pokud pozorované veličiny mají spojité rozložení, jde o test nulové hypotézy nezávislosti proti alternativní hypotéze o kladné závislosti. Hodnota 1 znamená kladnou závislost; jde o test na hladině α.

2. **Kendallův kvantifikátor** k-corr$_a$ pro $a \in (0, 0.5>$

$$\textbf{k-corr}_a(<t_1, t_2>) = 1 \quad \text{je-li} \quad \sum \text{sign}(R(i)-R(j))(\text{sign}(Q(i)-Q(j)) \geq k_a$$

pro $i \neq j$, kde k_a je vhodná konstanta. Statistická interpretace jako výše.

3. **Pořadově ekvivalenční kvantifikátor** e-corr$_a$

$$\textbf{e-corr}_a(<t_1, t_2>) = 1 \quad \text{je-li} \quad R(i) = Q(i) \quad \text{pro } i=1, \ldots, m$$

Korelační koeficienty se používají v sentencích podmíněného tvaru

F1 q F2 / F

kde F1 a F2 jsou reálné veli4iny a F je binární formule.

• Asociace základní

Pro testování spolehlivosti asociace se používají statistické charakteristiky, které jsou definovány pro binární data a vypočtené ze čtyřpolní tabulky shod a neshod antecedentů a sukcedentů.

Skutečná data obvykle nejsou jen binární. Z kapitoly o předzpracování dat víme, že je možno data kteréhokoliv datového typu převést na binární. Pokud máme data alespoň kategoriální, pak je obvykle pro asociace "binarizují" přímo dolovací algoritmy. Například tak, že testují postupně hodnoty atributu A=a1, A=a2, …, A=ak.

Zobecníme-li definici asociací, pak asociacemi nyní nazýváme vztahy mezi dvěma podmnožinami kategoriálních atributů (antecedentem a sukcedentem).

Je to tedy jeden z typů tvrzení

Jestliže $A=a \wedge B=b \wedge \ldots$ **pak** $X=x \wedge Y=y \wedge \ldots$ **se spol** S **a podp** P (1)

Hodnoty $A=a \wedge B=b \wedge \ldots$ **souvisí s** $X=x \wedge Y=y \wedge \ldots$ **se spol** S **a podp** P (2)

kde $\{A,B,\ldots\}$ jsou atributy antecedentu,

$\{X,Y,\ldots\}$ jsou atributy sukcedentu,

a, b, ..., x, y, … jsou hodnoty domén příslušných atributů,
S ... spolehlivost vztahu definovaná vhodnými kvantifikátory,
P ... podpora, počet případů, pro něž je nalezená hypotéza platná.

Výsledkem jsou takové vygenerované hypotézy, pro něž platí

$S \geq$ **Smin = minconf** (minconf je zadaná minimální **spolehlivost, confidence**)

P ≥ **Pmin = minsupp** (minsupp je zadaná minimální **podpora, support**)

❑ Metoda ASSOC

generuje zajímavé sentence pravdivé v binárních datech **X** tvaru

KONJ1 ~ KONJ2

kde ~ je některý asociační kvantifikátor
KONJ1, KONJ2 jsou elementární konjunkce.

Co považuje řešitel za zajímavé specifikuje zadáním vstupních parametrů:

a) Použitý kvantifikátor a jeho parametry (δ, α, Pmin, Smin)
Zvýšením parametrů δ, P, S a snížením parametru α se snižuje počet výstupních hypotéz (zvyšuje se nárok na teoretické hypotézy).

b) Povolený tvar antecedentu a sukcedentu, tj. zadá
maximální délku antecedentu a maximální délku sukcedentu,
antecedentové predikáty důležité BA a ostatní CA,
sukcedentové predikáty důležité BS a ostatní CS,
ke každému uvedenému atributu ATR jeho povolený tvar +ATR (jen pozitivní), -ATR (jen negativní), ATR (oba).

c) zda bude použito zlepšování, případně které (viz níže).

Poznámka:

- parametry z (a) a (b) definují množinu relevantních otázek RO,
- doplňující parametry δ, α, P, S množinu pravdivých sentencí H(X),
- bod (c) učí míru komprese výsledků.

❑ Metoda IMPL

generuje zajímavé sentence pravdivé v binárních datech **X** tvaru elementární implikace

KONJ ⇒* DISJ

kde ⇒* je některý implikační kvantifikátor

KONJ je elementární konjunkce

DISJ elementární disjunkce

Ostatní zadání metody je obdobné, jako u ASSOC.

❑ Algoritmy hledání asociací

Rozdíl mezi klasickými statistickými testy, prováděnými nad daty, a mezi dolováním asociací z dat je v automatizaci tohoto procesu. Dolovací algoritmy automaticky generují a testují "všechny možné" hypotézy a generují na výstupu ty, které projdou příslušným testem (s vhodným testovacím kritériem a zadaným hodnotami *minconf, minsupp*).

Problém všech algoritmů hledajících asociace je v testování "všech možných" sentencí, uvážíme-li počet teoretických případů, které je nutno otestovat, a jim odpovídající časovou složitost úlohy. Proto hlavním problémem celé metody je nalezení rychlých algoritmů, které by sice nalezly všechny požadované hypotézy, ale omezily by počet testů.

Algoritmy pro testování a generování hypotéz můžeme dělit na

- *triviální* - generují a testují všechny možné sentence jako kombinace hodnot atributů, délek ante a sukce, kombinací dvojic ante a sukce (odtud též kombinační analýza, pro větší data nepoužitelná, s exponenciální časovou složitostí),
- *uspořádané generování pravidel* – například vhodně uspořádané postupné prodlužování délky sukce a ante snižuje počet kladných shod a, pokud a < *minsupp*, negenerují se další sentence s větší délkou –cedentu,
- *vzorkováním* - rozsáhlá data zpracují po částech (vzorcích) a hledají kandidáty pro hypotézy, pak testují přes celá data jen kandidáty.

❑ Problém redundance výsledků

Při mechanickém generování asociací může nastat následující situace:

Platí-li $A=a \Rightarrow X=x$ se spolehlivostí $S = s$ (1)

a také $A=a \wedge B=b \Rightarrow X=x$ se spolehlivostí $S = s'$ (2)

Pak pro $s' = s$ říkáme, že (2) je redundandní (B=b je redundandní, nepřináší novou informaci)

$s' > s$ říkáme, že $B=b$ rozšiřuje sentenci (1)

$s' < s$ říkáme, že $B=b$ zužuje sentenci (1)

Za zajímavé obvykle považujeme rozšíření již nalezených hypotéz, naopak zúžení obvykle přiřadíme k redundancím.

Možnost rozšíření definice redundance: za redundandní považujeme i případy s' < s+C pro dané C.

Algoritmus N-ASOC (neredundandní asociace)

```
{DATA: array[1..r,1..s] of integer;
 SU: array[1..n,1..k] of integer…počet výskytů sukce
 ZP: array[1..r] of boolean … zpracovaný řádek dat
 m    = počet ante
 n    = počet sukce
 s=m+n
 k    = počet kategorií
 pod  = podpora, min. počet výskytů u hypotézy
 spol = spolehlivost, procento platných hypotéz
}

for u:=1 to r do
  begin
    if ZP[u]=0 then
      begin {nalezení všech stejných l-tic ante}
        for i:=u+1 to r do
          begin
            if ZP[i]=0 then
               if DATA[u]=DATA[i]
                 then begin a:=a+1; S[j,k]:=S[j,k]+1 end
          end;
      {zrusení redundancí, kdyz existuje vysledek s ante o 1 kratsi}
      {generování hypotéz, které projdou testem}
      end
  end;
```

ZP
|
|
|
…

A	B	C	…	X	Y	Z	…
1	2	3		1	3	3	
2	3	8		1	3	2	
1	2	3		2	3	1	
1	2	3		1	1	1	
2	3	8					
…							

S

	k1	k2	k3	…
X	||	|		
Y	|		||	
Z	||		|	
…				

Vlastnosti algoritmu N-ASOC:
- testuje jen kombinace hodnot skutečně přítomné v datech
- všechny sukce i všechny jejich délky jedním průchodem daty
- generuje jen neredundandní hypotézy pro danou definici redundance

Příklad 6.2.

Data o asistované reprodukci (AR) obsahují léčebné cykly AR, každý je popsán asi 250 atributy o pacientce, anamnéze, léčbě, dále časově dělené do několika etap, o dílčích výsledcích etap cyklu i výsledku celé léčby.

Jeden z vygenerovaných výsledků (použita FI), metoda AH = asistovaný hashing.

Je-li věk = 20-25 ∧ metoda AR = AH,
pak počet oplodněných oocytů > 0
s podp P = 112 **a spol** S = 76%

Interpretace: V datech existuje 112 případů pacientek věku 20-25 let, u kterých byla provedena metoda AH a z nich bylo 76% oocytů oplodněno.

Procento oplodněných vajíček v celém souboru AR je asi 50%. Na základě těchto výsledků můžeme formulovat hypotézu: U mladších pacientek věku 20-25 let metoda AH pozitivně ovlivňuje oplození vajíček.

♦

❑ Využití metod ASSOC a IMPL

Výsledky obou metod můžeme použít pro následující analýzy:

1. **analýza příčin** ... použijeme IMPL, setřídíme výsledky dle SUKCE
2. **analýza následků** ... použijeme IMPL, setřídíme výsledky dle ANTE
3. **analýza asociací** ... použijeme ASSOC, setřídíme výsledky podle potřeby
4. **konkrétní dotaz** ... použijeme metodu dle dotazu; při konkrétním dotazu máme možnosti:

   buď dotaz → analýza na míru → výsledek

   nebo provedení komplexní.analýzy → dotaz → výběr výsledku
5. **tvorba báze pravidel** ... použijeme IMPL, setřídíme výsledky, výsledná pravidla chápeme jako pravidla pro použití u dalších případů z reality

Příklad 6.3.

Matky, které daly své dítě k adopci si to někdy později rozmyslí. Úkolem je najít pravidla rozpoznávající, které atributy (příčiny) napovídají, zda si matka své rozhodnutí rozmyslí nebo ne. Data jsou od 104 matek se 23 atributy

Matka (m-věk, m-povolání, m-stav, m-národ, o-stav, o-vztah, o-povol, o-národ, počet-těhot,

počet-potrat, ... , kojení, ... , rozmyslela)

Řešení: Zvolena metoda IMPL, kvantifikátor FI, délka antecedentu 3

Výsledek – část úplného řešení:

Je-li	**pak**	**Spolehlivost**	**Podpora**
m-věk < 30	rozmyslela	66	62
m_věk ∈ <30, 40>	nerozmyslela	100	28
m_věk > 40	nerozmyslela	91	14
m_povol = dělnice, pomocná	nerozmyslela	100	19
o_povol = není ve vězení	nerozmyslela	90	86
o_vztah = ženatý jinde	nerozmyslela	100	67
poč_těhot = 3 ∧ poč_potrat = 0	nerozmyslela	100	22
útěk = ano	nerozmyslela	100	17
steril = ano	nerozmyslela	100	6
m_povol = prostituce	nerozmyslela	100	15
ústavní_výchova = ano	nerozmyslela	100	21
podvod s OP = ano	nerozmyslela	100	5
…			

Výsledek formulovaný slovně, uspořádaný podle síly pravidla:

Jestliže platí

m-věk>= 30 let nebo
m-povolání = dělnice,pom.síla,THP nebo
o-vztah=ženatý s jinou, rozvedený nebo
poč-těhot>=3 ^ poč-potratů=0 nebo
prostituce = ano

pak rozmyslela = ne
s podporou P = … **a spolehlivostí** = **100%**

Jestliže platí

m-věk <= 30 let ^ o-věk <= 30 let

pak rozmyslela = ne
s podporou P = ... **a spolehlivostí** = **66%**

Jestliže platí

m-věk <= 18 let

pak rozmyslela = ne
s podporou P = ... **a spolehlivostí** = **57%**

...

♦

Příklad 6.4.

Zadání: Japonská banka sleduje data o 125 lidech, žádajících banku o úvěr. Úkolem je z dosavadních „ručně“ prováděných rozhodnutí formulovat pravidla, podle kterých se bude v budoucnu rozhodovat automaticky. Evidence má atributy:

A1. dostal úvěr [A/N]
A2. je nezaměstnaný [A/N]
A3. čím ručí [1 = PC, 2 = car, 3 = stereo, 4 = bike]
A4. pohlaví
A5. svobodný [A/N]
A6. problematický region [A/N]
A7. věk [kateg]
A8. hotovost u banky
A9. měsíční příjem
A10. počet měsíců splátky
A11. počet roků pracujících v současné firmě

Řešení: Metoda IMPL,
parametry ANTE = {A2 .. A11}, SUKCE = A1
kvantifikátor = FI
min. spolehlivost S = 90%, min. podpora P = 1
délka ante = 3

Výsledky:

Je-li věk = (65, 81> **pak** úvěr nedostal s S =100%
Je-li hotovost = (40,50 > a současně měs.příjem = (6, 10> **pak** úvěr nedostal s S =100%
Je-li nezaměst = Ano a současně měs.příjem = (6, 10> **pak** úvěr nedostal s S =100% ...
Je-li roků u firmy = >3 a současně pohl = muž a současně počet splátek = (6, 20>
pak úvěr dostal s S = 96.7%

♦

❑ Dedukční pravidla

Ze sentencí pravdivých v datech je možno jistými úpravami odvozovat další pravdivé sentence. Pravidla pro takové odvozování nazýváme dedukčními pravidly. Pomocí dedukčních pravidel je možno odvozovat další pravidla a není nutné je vyhledávat v datech.

Platí:

F1 a F2 jsou ekvivalentní v datech, platí-li Oi(F1)=Oi (F2) , i=1,..., m

1. Pravidlo záměny ekvivalentních formulí.

V každé sentenci S(F) obsahující formuli F a pravdivé v datech X lze nahradit formuli F jinou formulí F’, která je v datech ekvivalentní formuli F. Vzniklá sentence je opět pravdivá v datech X.

Pro elementární konjunkce **K** a elementární disjunkce **D** platí zákony logiky: komutativní, negace-negace, De Morganovy ⇒ pro danou formuli tvaru **K** nebo **D** existuje řada **logicky ekvivalentních** formulí.

Příklad 6.5.

v datech X je pravdivá sentence

kouření $\Rightarrow_{0.8}$ infarkt $\vee$ rakovina_plic

pak jsou v datech pravdivé i sentence

kouření $\Rightarrow_{0.8}$ rakovina_plic $\vee$ infarkt
kouření $\Rightarrow_{0.8}$ $\neg$ $\neg$ (infarkt $\vee$ rakovina_plic)
kouření $\Rightarrow_{0.8}$ $\neg$ ($\neg$infarkt $\wedge$ $\neg$ rakovina_plic)

♦

Další ekvivalence mohou vyplynout z dat: 2 formule jsou **ekvivalentní v datech X**, mají-li pro všechny objekty shodné hodnoty: Oi(F) = Oj(F) pro všechna i, j.

Příklad 6.6.

v datech X platí

maj_auta je ekvivalentní s **maj_auta $\wedge$ muž**

pak ze sentence pravdivé v datech X (ne v realitě)

maj_auta $\Rightarrow_{0.9}$ má_garáž

pak i následující tvrzení jsou pravdivá v datech X (pravidlo 1)

maj_auta $\wedge$ muž $\Rightarrow_{0.9}$ má_garáž
muž $\wedge$ maj_auta $\Rightarrow_{0.9}$ má_garáž

♦

Praktické využití:

stačí testovat elementární konjunkce a disjunkce jednoho pořadí

2. Pravidlo úprav elementární implikace.

Elementární implikací je sentence tvaru K Þ* D kde K je elementární konjunkce, D je elementární disjunkce a K, D nemají žádný společný atribut. Platí: v elementární implikaci K Þ* D pravdivé v datech X můžeme

(1) převést člen z antecedentu do sukcedentu za současné změny znamení (formuli A nahradit její negací);

(2) přidat do sukcedentu nové členy.

Vzniklá sentence je opět pravdivá v datech.

Příklad 6.7.

v datech X je pravdivá sentence

otylost $\wedge$ muž $Þ_{0.8}$ infarkt

odtud plyne i pravdivost v datech pro sentence

otylost $\wedge$ muž $\Rightarrow_{0.8}$ infarkt $\vee$ ang.pect $\vee$ berc.vred
otylost $\Rightarrow_{0.8}$ $\neg$ muž $\vee$ infarkt $\vee$ ang.pect
muž $\Rightarrow_{0.8}$ $\neg$otylost $\vee$ infarkt

♦

Praktické využití:

- z pravdivosti jedné sentence můžeme rychle logicky odvodit (bez ověřování v datech) množství dalších sentencí pravdivých v datech
- při vhodném rozdělení atributů na ante – sukce je možno optimalizovat délku ante - sukce pro konkrétní algoritmus, skutečné rozdělení na ante – sukce pak použitím pravidla 2 upravit

3. Pravidlo symetrie.

Je-li $\sim$ * symetrický kvantifikátor a je-li sentence F1 $\sim$ * F2 pravdivá v datech X, pak i sentence F2 $\sim$ * F1 je pravdivá v datech.

Příklad 6.8.

v datech X je pravdivá sentence

$$\textbf{kouření} \wedge \textbf{víc\_30\_let} \Leftrightarrow_{0.8} \textbf{rakovina\_plic}$$

pak jsou v datech pravdivé i sentence

$$\textbf{rakovina\_plic} \Leftrightarrow_{0.8} \textbf{kouření} \wedge \textbf{víc\_30\_let}$$

♦

Praktické využití:

při testování symetrických asociací stačí jeden test

3. Pravidlo konzervativního zlepšování.

Řekneme, že atribut A (nebo $\neg$A) konzervativně zlepšuje elementární konjunkci K v datech X, jestliže A se nevyskytuje v K a formule K $\wedge$ A (nebo K $\wedge \neg$A) je ekvivalentní formuli K v X. V sentenci pravdivé v datech a obsahující elementární konjunkci K lze ke K přidat do konjunkce libovolný počet atributů a negovaných atributů, které K konzervativně zlepšují. Výsledná sentence je opět pravdivá v datech. Jde o speciální případ pravidla 1.

Příklad 6.9.

platí-li

$$\textbf{A1} \wedge \neg\textbf{A3} \wedge \textbf{A5} \Rightarrow \textbf{A4}$$

a zlepšují-li A5, $\neg$A7, A9 konzervativně antecedent, platí i

$$\textbf{A1} \wedge \neg\textbf{A3} \wedge \textbf{A5} \Rightarrow \textbf{A4}$$

$$\textbf{A1} \wedge \neg\textbf{A3} \wedge \textbf{A7} \Rightarrow \textbf{A4}$$

$$\textbf{. . .}$$

$$\textbf{A1} \wedge \neg\textbf{A3} \wedge \neg\textbf{A7} \wedge \textbf{A9} \Rightarrow \textbf{A4}$$

$$\textbf{A1} \wedge \neg\textbf{A3} \wedge \textbf{A5} \wedge \neg\textbf{A7} \wedge \textbf{A9} \Rightarrow \textbf{A4}$$

♦

Praktické využití: jako u pravidla 1

❑ Asociace s neúplnou informací

Dána datová matice **X** binární s chybějícími údaji, tedy hodnotami {0, x, 1}, kde x je neznámá hodnota.

Dvouhodnotovým doplněním X je jsou taková data **X'**, že se x nahradí 0 nebo 1. Zřejmě jen jediné doplnění je správné (odpovídá realitě), ale to neznáme. Proto bereme v úvahu všechna možná doplnění a podle všech jejich výsledků můžeme dělat závěry.

Platí **princip zabezpečení**: není.li výsledek jistý, platí hodnota x.

Jestliže A1, . . . , An jsou atributy antecedentu a sukcedentu, pak

K = O(A1 ∧ A2 ∧ . . . ∧ An) = 1 když A1 = A2 = . . . = An = 1
0 některé Ai = 0, ostatní Aj = 1
x některé Ai = x

D = O(A1 ∨ A2 ∨ . . . ∨ An) = 1 když některé Ai = 1
0 A1 = A2 = . . . = An = 0
x některé Ai = x

Chápeme-li 0 < x < 1, pak

$$O(A1 \wedge A2 \wedge \ldots \wedge An) = \min(O(A1), \ldots, O(An))$$

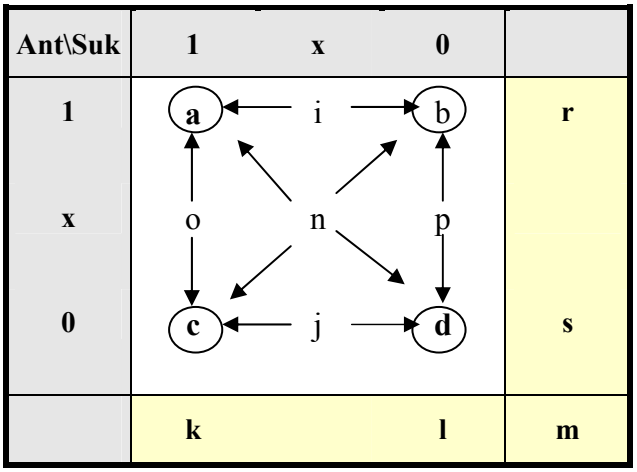
$$O(A1 \vee A2 \vee \ldots \vee An) = \max(O(A1), \ldots, O(An))$$

Princip zabezpečení říká, že elementární **K** nebo **D** má pro libovolný objekt hodnotu x právě tehdy, když existuje doplnění, že K(D) = 0 a jiné, že K(D) = 1.

Trojhodnotová frekvenční tabulka se převádí na dvojhodnotovou tak, že se

i se dílem přičte k **a**, dílem k **b**

o -"- **a** -"- **c** atd.



Příklad 6.10.

Data o 15 pacientech Pacient(..., kouření, otylost, vysoký_tlak) *s hodnotami NULL.*

Trojhodnotová tabulka:

antecedent: kouření ∧ otylost (*označíme* riziko)

testujeme implikaci ***riziko ⇒ tlak***

riziko \ tlak	1	x	0
1	3	0	0
x	2	0	0
0	1	0	9

pro tuto tabulku existují 3 možná doplnění:

(1)

5	0
1	9

(2)

4	0
2	9

(3)

3	0
3	9

*jsou pravdivá všechna (a/(a+b)=1) ⇒ sentence je pravdivá v datech (=**1**)*

♦

Příklad 6.11.

Data Pacient(..., kouření, otylost, vysoký_tlak) *s NULL.*

Trojhodnotová tabulka:

antecedent: kouření ∧ otylost *(označíme* riziko*)*

testujeme implikaci **¬riziko ⇒ ¬tlak**

¬riziko \ ¬tlak	1	x	0
1	9	0	1
x	0	0	2
0	0	0	3

pro tuto tabulku existují 3 možná doplnění:

(1)

9	1
0	5

(2)

9	2
0	4

(3)

9	3
0	3

je pravdivé 1. doplnění, nepravdivé 2. a 3. ⇒ sentence má v datech hodnotu **x.**

♦

Závěr: Data s větším počtem objektů a větším počtem neúplných hodnot v antecedentu a sukcedentu (tj. s velkými četnostmi i, o, n, p, j) jsou prakticky nezpracovatelná. Pro každou sentenci by se muselo testovat velké množství možných doplnění a výpočet by se velmi prodražil.

Shrnutí pojmů 6.2.

Základní asociace. Formule. Kvantifikátory. Antecedent a sukcedent. Sentence, hypotéza.

Zobecněná asociace a její tvary. Spolehlivost a podpora hypotézy.

Dedukční pravidla.

Asociace pro neúplná data.

Otázky 6.2.

1. Vysvětlete podstatu jednoduchých asociací.
2. Z čeho vychází základní princip hledání asociací?
3. Jak se liší metoda hledání asociací od testování hypotéz o závislosti dvou veličin?
4. K čemu jsou prakticky užitečná dedukční pravidla pro asociace?
5. Je možno prakticky na velikých datech s neúplnou informací nalézt asociace?

Úlohy k řešení 6.2.

1. Jsou dána data z kapitoly 5, úlohy 1 až 4. Navrhněte pro ně využití hledání asociací: zadejte antecedenty, sukcedenty, kvantifikátor a hodnoty minimální spolehlivosti a podpory. Všechny volby v zadání zdůvodněte a formulujte, co očekáváte jako možný výsledek.

6.3. Asociace transakční

Cíl Po prostudování této kapitoly budete

- vědět, co jsou základní asociace mezi podmnožinami atributů,
- umět popsat jejich význam, interpretaci,
- pro praktické úlohy navrhnout smysluplné typy úloh na asociace

Výklad

Jiný častý tvar zdrojových dat pro asociace je tzv. nákupní košík. Objektem je jeden (obvykle obchodní) případ, jeho několik atributů má obvykle pevnou strukturu (datum, čas, zákazník, … = identifikace košíku). Vysoký počet dalších, obvykle binárních atributů (seznam nakupovaného zboží = obsah košíku) je zadáván jako seznam atributů nabývajících nenulové hodnoty.

Asociacemi se zde rozumějí nalezené podmnožiny atributů, vyskytujících se společně (v košíku).

A	B	…	X – seznam (i >> 1]
a1	b1	…	X1 = {x1,x2,x3,x4,x5}
a2	b2	…	X2 = {x3,x8}
a3	b3	…	…
…			

Asociací zde rozumíme jeden z typů tvrzení

$$\{x1, x2, \ldots\} \in Xi \textbf{ se spol } V \textbf{ a podp } P \qquad (3)$$

$$\text{Je-li } A=a \wedge B=b \wedge \ldots \text{ pak } \{x1, x2, \ldots\} \in Xi \text{ se spol } V \text{ a podp } P \qquad (4)$$

Příklad 6.12.

transakčních asociací: Klasickým příkladem je vydolování "znalosti" z databáze o prodeji v supermarketu typu

"{jogurt, vločky} se spol 72% a podp 4567"

nebo

"je-li pátek ∧ léto, pak {pivo, buřty} se spol 67% a podp 3367"

V prvním případě se uvedená zboží umístí v prostoru co nejdále od sebe, ve druhém se v pátek vedle piva a buřtů umístí další víkendová lákadla.

♦

Algoritmy hledání transakčních asociací

Pro transakční asociace je potřeba použít jiné algoritmy, než pro asociace klasické. Jsou založeny na různých principech, využívají indexování, konstrukce tzv. k-množin apod., viz [x].

Shrnutí pojmů 6.3.

Asociace transakční jako podmnožiny společně se vyskytujících atributů.

Nákupní košík, jeho identifikace a obsah.

Interpretace a využití výsledků transakčních asociací.

Otázky 6.3.

1. Vysvětlete podstatu transakčních asociací a jejich rozdíl od jednoduchých asociací.
2. Pro která data se používají transakční asociace?
3. Je možno použít stejné algoritmy pro jednoduché a pro transakční asociace? Proč?

Úlohy k řešení 6.3.

1. Najděte alespoň 5 příkladů z praxe pro využití transakčních asociací, pro každý formulujte úplné zadání.

6.4. Asociace agregované

Cíl Po prostudování této kapitoly budete

- vědět, co jsou základní asociace mezi podmnožinami atributů,
- umět popsat jejich význam, interpretaci,
- pro praktické úlohy navrhnout smysluplné typy úloh na asociace

Výklad

Agregovanou asociací nazveme neprůměrný vztah mezi hodnotami množiny antecedentových atributů {A, B, C…} a skupinovými ukazateli {U1, U2, …} vypočtenými z atributů sukcedentu {X, Y, …} tvaru

Je-li A=a ∧B=b ∧… , **pak** U1=u1(**V1**) ∧U2=u2 (**V2**) ∧… **s podp** *P* (5)

kde A, B, … , a, b, … mají stejný význam jako u klasických asociací
U1, U2, … jsou identifikátory definovaných agregovaných ukazatelů,
u1, u2, … vypočtené hodnoty ukazatelů pro testovanou skupinu,
V1, V2, … numericky nebo symbolicky označují, zda je Ui průměrný, statisticky významně podprůměrný nebo nadprůměrný vzhledem k základnímu souboru
P je počet výskytů skupiny neboli podpora hypotézy.

Příklad 6.13.

Porovnání výsledků asociací klasických a agregovaných

Opět použijeme data o asistované reprodukci. Skupina vygenerovaných výsledků klasickými asociacemi (použita FI, TR = transfer embryí do dělohy) je:

Jestliže		**pak**	**se spol**	**s podporou**
TR provedl gynekol =	A	výsledek gravid = netěh	82%	3251
	B	netěh	72%	2865
	C	netěh	79%	1233
	D	netěh	85%	865

Takto formulované výsledy nejsou zajímavé, ale tyto negativní výsledky (binárních hodnot) je možno interpretovat jako pozitivní:

Jestliže		**pak**	**se spol**	**s podporou**
TR provedl gynekol =	A	výsledek gravid = těh	18%	3251
	B	těh	28%	2865
	C	těh	21%	1233
	D	těh	15%	865
	…			

Průměrné procento otěhotnění po transferu je 17%. Vzhledem k různému počtu případů u různých lékařů není možné bez statistického testu rozhodnout, zda a který lékař dosahuje nad- či podprůměrných výsledků a je-li takový výsledek zajímavý.

Vygenerovaný výsledek metodou agregovaných asociací. Použijeme ukazatel GR*100/ET, *kde GR = počet gravidit, ET = počet provedených embryotransferů, průměrný GR*100/ET za celý soubor je 17%:*

Jestliže		**pak**		**s podporou**
TR provedl gynekol =	A	GR/ET = 18%	-	3251
	B	28%		2865
	C	21%	.	1233
	D	15%	++	865
	…			

♦

❑ Test významnosti generovaných hypotéz

Pro otestování významnosti (neprůměrnosti) generovaných hypotéz použijeme následující výpočet pomocí hodnoty p-value.

Označíme p ... průměrná hodnota ukazatele za celý soubor
ps ... průměrná hodnota ukazatele za skupinu
ns ... počet případů ve skupině

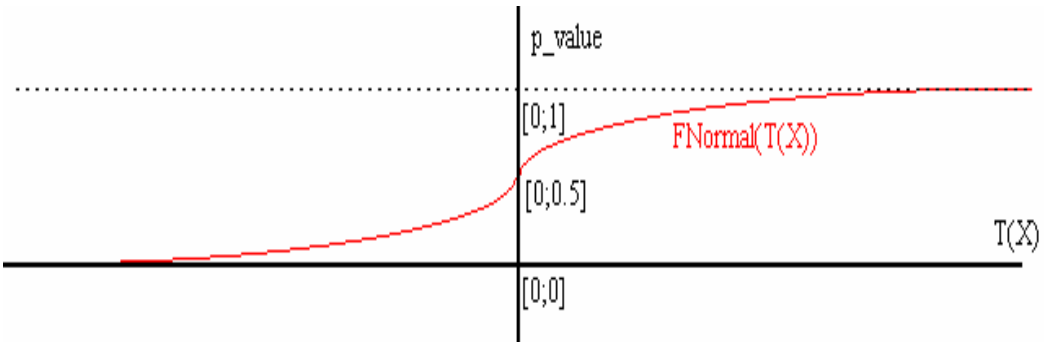
Testovací statistika

$$T(X) = \frac{ps - p}{\sqrt{p * (1 - p)}} * \sqrt{ns}$$

$$P(X) = FNormal(T(X))$$

$$p - value = 1 - P(X)$$

Průběh normální distribuční funkce má následující graf a jeho hodnotu pro vypočítanou hodnotu T(X) dostaneme standardní procedurou FNormal.



Obr.x. Průběh hodnoty FNormal (T(X))

Výsledné rozhodnutí o statistické významnosti výsledné hypotézy zapíšeme pro uživatele snadno čitelnou jednoduchou formou. Místo hodnoty p-value, kterou by si musel inpertpretovat do slovního vyjádření podle následující tabulky:

Hodnota p-value	+podm	označení	výsledek
p-value > 0.05		=	průměr
p-value ∈ <0.05, 0.01>		.	???
p-value < 0.01 (0.001)	ps > p	+, ++	nadprůměr
p-value < 0.01 (0.001)	ps < p	-, --	podprůměr

❑ Algoritmus výpočtu asociací agregovaných

Vlastní výpočet agregovaných ukazatelů není náročnější, než výpočet kvantifikátoru u klasických asociací, proto je opět určující postup generování kombinací antecedetových atributů. Protože se opět počítají všechny ukazatele současně, je výpočet časově srovnatelný s klasickými agregacemi.

Algoritmus AGR-ASOC

1. Definují se atributy antecedentu, pro sukcedent potřebné proměnné, ukazatele a podmínky pro proměnné.
2. První průchod daty vypočte proměnné a ukazatele za celá data.
3. Generují se sentence jako kombinace hodnot antecedentu (1 kombinace = 1 skupina), k nim hodnoty proměnných a ukazatelů.

   Proměnná je definovaná agregovaná hodnota (počet, suma, ...), z níž se počítají ukazatele.
4. Každý ukazatel se otestuje na statistickou významnost pomocí hodnoty p-value.
5. Všechny výsledky nebo jen významné se uloží.

Vlastnosti algoritmu AGR-ASOC

- testuje jen kombinace hodnot skutečně přítomné v datech
- všechny ukazatele skupiny testuje jedním průchodem daty
- generuje všechny neprůměrnosti daných ukazatelů

Příklad 6.14.

Jsou dána data o asistované reprodukci o 8500 záznamech léčebných cyklů a asi 120 atributech, mj. atributech o počtech předcházejících porodů, potratů, ...o počtech odebraných a oplozených vajíček apod. Z nich expert určí charakteristické ukazatele, vypočtené ze sum a počtů případů (proměnných):

Proměnné		*podmínka*
CYKL	počet léčebných cyklů AR	-
AS	počet aspirací	TC16 = 0
OO	suma odebraných oocytů OO21	TC16 = 0
OPL	suma oplozených oocytů TR26	TR26 > 0
ET	počet embryotransferů	TR29 > 0
GR	počet klinických gravidit po AR	GR > 0
POR	počet porodů po AR	GR = 9
AB	počet potratů po AR	GR = 1,2,3,4
EU	počet mimoděložních těhotenství po AR	GR = 5,6
VIC	počet vícečetných gravidit po AR	VGR = 1

Ukazatele

OO/AS	průměrný počet získaných oocytů při aspiraci
OPL/OO*100	procento oplozených oocytů ze získaných oocytů - Fertilisation rate FR
GR/ET *100	procento klinických gravidity na embryotransfer - Pregnancy rate PR
POR/GR*100	procento gravidit po AR ukončených porodem - Baby take home rate BR
AB/GR*100	procento gravidit po AR ukončených potratem
EU/GR*100	procento mimoděložních gravidit po AR
VIC/GR*100	procento vícečetných gravidit po AR

I. Jeden z úplných výsledků, kde jsou i statisticky nevýznamné výsledky pro porovnání vlivu všech hodnot antecedentu:

Věk matky

vek_k	**cykl**	**as**	**oo**	**opl**	**et**	**gr**	**por**	**ab**	**eu**
nevyplněno	8	8	13	4	3	0	0	0	0
<=25	1098	1037	5184	2552	484	86	43	34	9
26-30	3207	2976	13537	7046	1491	259	150	98	11
31-35	3069	2779	10850	5439	1455	237	129	91	17
36-40	1028	886	2792	1459	452	80	40	38	2
>=41	102	77	207	115	40	4	1	3	0
	8512	7763	32583	16615	3925	666	363	264	39

oo/as	**opl/oo**	**gr/et**	**por/gr**	**ab/gr**	**eu/gr**	**vic/gr**
1,62 =	30,77 =	0 =	0 =	0 =	0 =	0 =
5 =	49,23 =	17,77 =	50 =	39,53 =	10,47 =	12,79 =
4,55 =	52,05 =	17,37 =	57,92 =	37,84 =	4,25 =	15,83 =
3,9 =	50,13 =	16,29 =	54,43 =	38,4 =	7,17 =	16,03 =
3,15 =	52,26 =	17,7 =	50 =	47,5 =	2,5 =	18,75 =
2,69 =	55,56 =	10 =	25 =	75 =	0 =	0 =
50,99		16,97	54,5	39,64	5,86	15,77

Tento výsledek je velmi neočekávaný. Říká, že věk matky nemá žádný podstatný vliv na úspěšný výsledek léčby (nikde se nevyskytují nad- nebo podprůměrné hodnoty ukazatelů). Přitom všeobecně se předpokládá opak.

II. Výsledky jen signifikantní, statisticky významné pro jeden z antecedentů:

Je-li EU1 = 0	pak	FR	= 49.2 - -	s podp 6985
= 0		EU/GR	= 0.8 - -	6985
= 1		FR	= 56.9 + +	985
= 1		EU/GR	= 25.5 + +	985
= 2		FR	= 59.4 + +	472
= 2		POR/GR	= 32.7 -	472
= 2		EU/GR	= 14.6 +	472
= 3		FR	= 59.2 +	52
= 3		EU/GR	= 50.0 + +	

Závěr: *Pacientky bez EU mají nižší riziko EU a nižší FR,*
pacientky po EU mají vyšší riziko EU a vyšší FR,
vyšší počet předchozích EU zvyšuje riziko EU po AR.

♦

Shrnutí pojmů 6.4.

Asociace agregované, jejich tvar a rozdíl proti asociacím klasickým.

Ukazatele charakterizující agregační skupiny.

Otázky 6.4.

1. Jaký je rozdíl mezi klasickými asociacemi a asociacemi agregovanými?
2. Najděte v rámci látky tohoto předmětu metodu nejpodobnější výsledkům agregovaných asociací.
3. Jak by se daly výhodně zobrazit výsledky agregovaných asociací?

Úlohy k řešení 6.4.

1. Jsou opět dána data z kapitoly 5, úlohy 1 až 4. Navrhněte pro ně využití hledání agregovaných asociací: zadejte antecedenty, ukazatele, kvantifikátor a hodnoty minimální spolehlivosti a podpory. Všechny volby v zadání zdůvodněte a formulujte, co očekáváte jako možný výsledek.