STUDIUM ZÁVISLOSTI V KONTINGENČNÍCH TABULKÁCH

Dan Pokorný, Ulm

Práce je věnována památce načetínských lesů.

Tento příspěvek se zabývá problematikou studia závislosti dvou diskrétních, zejména nominálních, znaků. Zájem je věnován úloze prvotního zjištění a účelné representace struktury závislostí, pozornost je věnována kolapsování dvourozměrných tabulek.

V první kapitole jsou popsány základní míry a matice používané ke sledování souvislosti v dvourozměrných kontingenčních tabulkách.

Kapitola druhá se zabývá některými výběrovými vlastnostmi měr. Většina tvrzení zde obsažených byla zkoumána v kontextu logiky automatizovaného výzkumu, specielně jsou relevantní vzhledem k proceduře popsané v poslední kapitole. Obecně pak jsou tvrzení tohoto typu důležitá pro explorativní analýzu dat.

V kapitole třetí je pak podrobněji popsána jedna procedura pro zkoumání struktury závislosti v dvourozměrné kontingenční tabulce.

1. Základní postupy zkoumání závislosti v dvourozměrných kontingenčních tabulkách

1.1. Některé míry závislosti v RxC tabulkách

Uvažujeme dvourozměrné multinomiální rozdělení diskrétních veličin F a G, veličina F může nabývat hodnot z množiny R = {1,...,R}, veličina G

$$
\begin{array}{r|ccccc|l}
 & G= & & & & & \\
 & 1 & \dots & j & \dots & C & \\
\hline
F=1 & n_{11} & \dots & n_{1j} & \dots & n_{1C} & n_{1.} \\
 & \vdots & & & & & \\
i & n_{i1} & \dots & n_{ij} & \dots & n_{jC} & n_{i.} \\
 & \vdots & & & & & \\
R & n_{R1} & \dots & n_{Rj} & \dots & n_{RC} & n_{R.} \\
\hline
 & n_{.1} & \dots & n_{.j} & \dots & n_{.C} & m=n_{..}
\end{array}
\qquad (T\ 1)
$$

hodnot z množiny C = {1,...C}. Pravděpodobnost, že veličina F nabyde hodnoty $i \in R$ označme $p_{i.}$ obdobně $p_{.i}= P(G=j)$ a $p_{ij}= P(F=i, G=j)$. Hypotézu o nezávislosti veličin F a G: $(\forall\ i,j)\ (p_{ij}=p_{i.}p_{.j})$ proti alternativě závislosti $(\exists\ i,j)\ (p_{ij}\neq p_{i.}p_{.j})$ můžeme testovat, pokud máme k dispozici konečný výběr o rozsahu m z uvažovaného rozdělení. Označme n_{ij} počet objektů z tohoto výběru, pro něž veličina F nabývá hodnoty i a současně veličina G hodnoty j, obvyklým způsobem označíme i marginální četnosti; viz Tabulka 1.

Matici $(n_{ij})_{i=1,j=1}^{i=R,j=C}$ nazýváme dvourozměrnou kontingenční tabulkou. Nejobvyklejším testem nezávislosti je Pearsonův χ^2-test. Označme "očekávané četnosti" $e_{ij}=n_{i.}n_{.j}/m$. Hypotézu nezávislosti zamítáme., překročí-li statistika

$$CHISQ = \sum_i \sum_j (n_{ij} - e_{ij})^2 / e_{ij}$$

kritickou hladinu χ^2-rozdělení s (R-1)(C-1) stupni volnosti. Variantou Pearsonova testu je test založený na statistice

$$MLCHISQ = 2 \Sigma \Sigma\, n_{ij} \ln n_{ij}/ e_{ij}$$

(Pozn.: klademe 0 ln 0 = 0.)

Poněkud komplikovanější je statistika založená na maximálně významné logaritmické interakci. Matici RxC reálných čísel (a_{ij}) nazveme interakční maticí, splňuje-li podmínky:

(a) všechny řádkové součty jsou nulové,
(b) všechny sloupcové součty jsou nulové,
(c) existuje $a_{ij} \neq 0$.

Testovou statistikou je nyní výraz

$$WSQ = \max((\Sigma\Sigma\, a_{ij} \ln n_{ij})^2 / (\Sigma\Sigma\, a_{ij}^2 / n_{ij})),$$

kde maximum je uvažováno přes všechny možné interakční matice (a_{ij}). Nulové četnosti n_{ij} je doporučeno nahradit hodnotou 0.5.

Statistiky MLCHISQ i WSQ opět srovnáváme s χ^2-rozdělením s (R-1)(C-1) stupni volnosti.

Kromě uvedených testů nezávislosti se užívají dále míry souvislosti vyjadřující stupeň těsnosti vztahu mezi veličinami. Nejužívanější z nich jsou odvozeny ze statistiky CHISQ:

$$PSISQ = CHISQ / m$$

nebo "Cramerovo V"

$$VSQ = CHISQ/(m(\min(R,C)-1))$$

nabývající hodnot mezi 0 a 1.

Existuje ovšem celá řada dalších měr a statistik pro RxC tabulky, užitelných ve speciálních situacích (např. jeden či oba znaky jsou ordinální, vyšetřujeme predikovatelnost jednoho znaku z druhého, zkoumáme apriorně daný typ souvislosti a pod.).

1.2. Některé míry souvislosti v 2x2 tabulkách

Zjednodušíme si označení z (T1) a četnosti budeme značit

	G = 1	2	
F = 1	a	b	r
2	c	d	s
	k	l	m

(T2)

Analogiemi statistik CHISQ, MLCHISQ a WSQ budou následující statistiky CHI, MLCHI a W, které však nyní v testech nezávislosti $(p_{11}=p_{1.}p_{.1})$ proti alternativě kladné asociovanosti $(p_{11}>p_{1.}p_{.1})$ budeme srovnávat s kvantily normálního rozdělení N(0,1).

$$CHI = (ad-bc)\sqrt{(m)}/\sqrt{(klrs)} = (am-kr)\sqrt{(m)}/\sqrt{(klrs)}$$

$$MLCHI = \sqrt{(2.(a \ln a + b \ln b + c \ln c + d \ln d - k \ln k - l \ln l - r \ln r - s \ln s + m \ln m))} \quad \ldots \text{ pro } ad-bc>0$$
$$= 0 \quad \ldots \text{ jinak}$$

$$W = \ln((ad)/(bc)/ \quad (1/a+1/b+1/c+1/d \quad \ldots \text{pro } ad-bc>0$$
$$= 0 \quad \ldots \text{ jinak}$$

Místo b=0 resp. c=0 klademe opět b=0.5 resp. c=0.5, avšak pro a=0 nebo d=0 položíme nyní W=0.

Známou modifikací CHI statistiky je Yatesova korekce:

$$YATES = (ad-bc-m/2)\sqrt{(m)}/\sqrt{(klrs)} \ldots \text{ pro } (ad-bc)>m/2$$
$$= 0 \quad \ldots \text{ jinak}$$

Všechny dosud uvedené testy jsou asymptotického charakteru. Přímo odvozený Fisherův test souvislosti je dopoporučován zejména pro výběry malého rozsahu, přesněji pro tabulky, v nichž alespoň některá z očekávaných četností e_{11}, e_{12}, e_{21}, e_{22} je malá.

$$FISHER = \sum_{i=0}^{\min(k,r)-a} (k!\,l!\,r!\,s!)/((a+i)!\,(b-i)!\,(c-i)!\,(d+i)!\,m!)$$

Alternativní hypotézu souvislosti na hladině α přijmeme pokud FISHER $\le \alpha$.

Poznámka: Naznačíme algoritmus pro výpočet Fisherova testu, který (a) je velmi efektivní a (b) nezpůsobí případnou chybu "overflow" při výpočtu faktoriálů, resp. kombinačních čísel:

(1) Před výpočtem Fisherova testu pro sérii čtyřpolních tabulek (T2) vytvoříme pomocné pole F(0:MMAX), kde MMAX≥m pro všechny tabulky. V poli F uložíme logaritmy faktoriálů rekursivně vypočtené:

$$F(n) = \ln(n) + F(n-1).$$

Pro danou tabulku (T2):

(2) Vypočteme $Q = F(k) + F(l) + F(r) + F(s) - F(m)$

(3) Vypočteme

$$FISHER = \sum_{i=0}^{\min(k,r)-a} \exp(Q-F(a+i)-F(b-i)-F(c-i)-F(d+i))$$

Analogií míry PSISQ (a zároveň míry VSQ) je

$$PSI = CHI/\sqrt{(m)}$$

Dalšími mírami jsou interakce, logaritmická interakce a Yuleovo Q:

INTERACTION = (ad) / (bc)

LOGINT = ln (INTERACTION)

YULE = (ad-bc) / (ad+bc)

LOGINT i YULE jsou (a) funkčně závislé na INTERACTION, všechny tři míry jsou (b) invariantní vůči přenásobení řádků či sloupců tabulky (T2) kladnými konstantami. (a) a (b) jsou ekvivalentní podmínky.

Zobecnění těchto měr pro případ RxC tabulky je možné, ale není jednoznačné.

1.3. Interpretace souvislosti v RxC tabulce pomocí RxC matice hodnot

Uvedené testy a míry nedávají prakticky žádnou informaci o charakteru případné závislosti veličin. Lepší přehled mohou poskytnout různé RxC matice hodnot, informující o významu jednotlivých políček tabulky (T1).

Takovými maticemi mohou například být:

(1) matice pozorovaných četností (n_{ij})

(2), (3), (4) Matice relativních řádkových (resp. sloupcových, resp. celkových) četností $(n_{ij}/n_{i.})$ (resp. $(n_{ij}/n_{.j})$, resp. (n_{ij}/m)).

(5) Mostellerovy "smoth values": Pokusíme se nalézt dva vektory kladných čísel $x=(x_1,\ldots,x_R)$ a $y=(y_1,\ldots,y_C)$ tak, aby v matici $S=(s_{ij})=(x_i y_j n_{ij})$ platilo

$s_{1.} = \ldots = s_{R.}$ a $s_{.1} = \ldots = s_{.C}$.

(6) Matice rozdílů pozorovaných a očekávaných četností $(n_{ij}-e_{ij})$

(7) Matice adjustovaných odchylek $(n_{ij}-e_{ij})/\sqrt{(e_{ij})}$

(8) Matice standardizovaných adjustovaných odchylek (Habermanovy odchylky): $(n_{ij}m - n_{i.}n_{.j})\sqrt{(m)}/\sqrt{(n_{i.}(m-n_{i.})n_{.j}(m-n_{.j}))}$

(9) Matice nejvýznamnější logaritmické interakce: interakční matice (a_{ij}), pro niž statistika WSQ nabývá svého maxima.

1.4. Kolapsování kontingenčních tabulek

Kolapsováním RxC kontingenční tabulky rozumíme vytvoření nové tabulky rxc (kde r≤R a c≤C), která vznikne vektorovým sečtením některých řádků nebo sloupců. Kolapsování tedy znamená sloučení některých kategorií sledovaných znaků:

R kategorii prvního znaku je sloučeno do r kategorií,
C kategorií druhého znaku do c kategorií.

Proč se kontingenční tabulky kolapsují?

(a) Pokud byly kategorie navrženy příliš podrobně, resp. jsou ve výběrovém vzorku zastoupeny velmi nerovnoměrně, pak některé z očekávaných četností e_{ij} mohou být "malé" (hranice "malosti" bývá nejčastěji udávána 5,3 nebo 1), v kterémžto případě se test nezávislosti pomocí CHISQ příliš nedoporučuje.

(b) I při dostatečných očekávaných četnostech e_{ij} může být neúčelně jemná kategorizace na závadu; zpravidla ztrácíme na síle testu nezávislosti.

(c) I v případě, že test nezávislosti mezi příliš jemně kategorizovanými znaky významně zamítl nulovou hypotézu, stále ještě můžeme tratit, RxC matice popsané v předchozím odstavci jsou příliš rozsáhlé ergo nepřehledné.

Všechny lepší statistické programové systémy /BMDP, SPSS a j./ proto obsahují prostředky pro snadné kolapsování kontingenčních tabulek.

Systém BMDP navíc umožňuje automatické kolapsování v případě ordinálních znaků: Uživatel zadá požadované MINIMUM očekávaných četností. V tabulce je nalezena nejmenší očekávaná četnost e_{ij} a buď i-tý řádek nebo j-tý sloupec je zkolapsován s jedním řádkem /sloupcem/ sousedním. Z nejvýše čtyř možných kandidátů je vybrán řádek /sloupec/ s nejmenší marginální četností. Postup se iteruje, až je získána kolapsovaná tabulka s očekávanými četnostmi rovnými nebo většími zadanému číslu MINIMUM.

Spotřebitelem navržené kolapsování není však jedinou obranou proti nepřehlednosti RxC matice hodnot; je možno pokusit se representovat strukturu závislosti v kontingenční tabulce úspornějším způsobem. Zde je k dispozici řada postupů, viz [3],[5],[7] a [11]. V poslední kapitole popíšeme proceduru využívající k interpretaci struktury souvislosti automatického, "numericky optimálního", kolapsování dvourozměrné tabulky.

2. Výběrové vlastnosti měr v dvourozměrných kontingenčních tabulkách

Pro vývoj procedur, užívajících opakované míry /a statistiky/ pro dvourozměrné kontingenční tabulky, se ukázala vhodnou dobrá znalost chování těchto měr na úrovni výběrových souborů.

V této kapitole uvádíme některé výsledky tohoto typu, které by mohly být obecněji zajímavé.

2.1. Asociativnost měr v 2x2 tabulkách

Def. Prostorem čtyřpolních tabulek nazveme množinu

$$T = \{ \left(\begin{smallmatrix} a b \\ c d \end{smallmatrix}\right) ; \quad a, b, c, d \quad \text{celá nezáporná} \}$$

Def. Relace ("je asociativně lepší") na prostoru čtyřpolních tabulek je definována vztahem:

$\left(\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right) \geq \left(\begin{smallmatrix} a b \\ c d \end{smallmatrix}\right)$ právě když $A \geq a$, $B \leq b$, $C \leq c$, $D \geq d$.

Def. Míra pro čtyřpolní tabulky je zobrazení z prostoru čtyřpolních tabulek do množiny reálných čísel, rozšířené o prvky +/- ∞.

Příklad: Mírami pro čtyřpolní tabulku jsou např. CHI, YATES, FISHER, YULE atd., ale též např. výsledek testu "CHI=1.69?", jehož hodnotou je 1 pokud byla přijata alternativní hypotéza (tj. pokud bylo pro danou tabulku CHI = 1.69) a 0 v případě opačném.

Def. Míra M je asociativní právě když je neklesající vzhledem k uspořádání ≥.

Asociativnost je dosti přirozeným požadavkem na míry resp. testy v 2x2 tabulce, chceme, abychom pro "lepší" tabulku dostali "lepší" výsledek: Míra M je asociativní právě když $A \geq a$, $D \geq d$ /"více pozorování v políčkách hlavní diagonály"/ a $B \leq b$, $C \leq c$ /"méně v políčkách vedlejší diagonály"/ implikuje $M(A,B,C,D) \geq M(a,b,c,d)$.

Tvrzení: Míry CHI, PSI, INTERACTION, LOGINT, YULE a /1-FISHER/ jsou asociativní. Test nezávislosti proti kladné souvislosti, založený na CHI i test Fisherův jsou asociativní pro každou hladinu významnosti.
Důkazy: viz [8], str.163.

Tvrzení: Míra W n e n í asociativní.

Test nezávislosti proti kladné souvislosti založený na W je asociativní pro hladiny významnosti $\alpha = 0.05$.
Důkazy: viz [13], resp.[8] str.358.

Poznamenejme, že pojem asociativnosti, navržený a poprvé zkoumaný Petrem Hájkem, hraje důležitou roli v metodách mechanizované formace hypotéz. V rámci této teorie byly navrženy metody analýzy "mnohorozměrných" kontingenčních tabulek pomocí vyšetřování hodnot jisté míry v mnoha odvozených 2x2 tabulkách /srovnej "asociační resp. implikační GUHA proceduru" v [8]/. Touto mírou v 2x2 tabulce může být právě jakákoli asociativní míra.

2.2. Poznámka o maximálních hodnotách měr v 2x2 tabulkách

Tvrzení předešlého odstavce o asociativnosti měr CHI, PSI, INTERACTION, LOGINT a YULE, tedy o jejich monotonii vůči uspořádání prostoru čtyřpolních tabulek ≥, doplníme nyní informací o jejich maximálních nabývaných hodnotách.

Tvrzení: INTERACTION nabývá své maximální možné hodnoty (+∞) pro tabulky (T2) právě když a.d>0 a b.c=0. Totéž platí i pro monotonní transformace LOGINT (+∞) a YULE (+1).

Tvrzení: Míra PSI nabývá své maximální možné hodnoty +1 právě když a.d>0 a b=c=0.

Mějme dva dvouhodnotové znaky, oba nekonstantní ve výběrovém souboru. Míra INTERACTION /LOGINT, YULE/ nabývá své maximální možné hodnoty v případě, že ve výběrovém souboru jeden znak logicky implikuje druhý nebo naopak.
Míra PSI /resp. CHI při fixovaném součtu m=a+b+c+d/ pak nabývá své maximální možné hodnoty právě kduž znaky jsou ve výběrovém souboru logicky ekvivalentní.

2.3. Srovnání měr a síly testu

V tomto odstavci srovnáme hodnoty některých měr a síly některých testů. Pro zkoumané míry uvedeme několik nerovností platících deterministicky, t.j. v každém výběrovém souboru. Pokud tyto míry budou statistikami, užívanými v testech se stejnými kritickými obory <K, +∞), získáme tím tvrzení o síle těchto testů.

Tvrzení: Pro každou 2x2 tabulku platí následující nerovnosti:
|PSI| ≤ |YULE|

YAYES ≤ CHI
W ≤ CHI

a rovnost mezi PSI a YULE nastává jen když PSI = 0 nebo |PSI| =1, rovnost mezi W a CHI jen když CHI = 0.
Důkaz: viz [14].

Tvrzení: Pro každou RxC tabulku kladných četností platí
WSO ≤ CHISQ
Důkaz: v rukopise.

Nerovnost posledních tvrzení může "pokažena" jen nahražením některé nulové četnosti n_{ij} hodnotou 0.5 pro WSQ statistiku. Nerovnost W ≤ CHI však platí vždy, viz definici W v odstavci 1.2. Pro RxC tabulky s nulovými resp. malými četnostmi někteří autoři doporučují přičítat malé číslo (0.5) ke každé pozorované četnosti; v tom případě by platila nerovnost WSQ ≤ CHISQ bez vyjímky.

Z uvedených nerovností plyne:

Důsledek: Test nezávislosti v 2x2 tabulce založený na Pearsonově statistice CHI je stejnoměrně silnější než test založený na logaritmické interakci. Obdobně pro Pearsonův a Yatesův test v 2x2 tabulce a - s uvedenými výhradami - pro Pearsonův a interakční test v RxC tabulce.

Znalost obdobných vztahů má význam nejen pro interpretaci výsledků výpočtů v konkrétních datech, ale i pro konstrukci efektivních perocedur: Chceme např. zjišťovat, zamítá-li test založený na interakční statistice WSQ hypotézu o nezávislosti ("WSQ ≥ CRIT ?"). Výpočetní čas můžeme uspořit, zjistíme-li nejdříve, zda CHISQ ≥ CRIT; pokud ne, pak časově náročnější výpočet WSQ vůbec neprovádíme.

2.4. Chování Pearsonovy statistiky vzhledem ke kolapsování kontingenční tabulky

Tvrzení: Buď S_1 hodnota statistiky CHISQ v tabulce RxC, S_2 hodnota statistiky CHISQ v tabulce (R-1)xC, která vznikla z předchozí kolapsováním dvou řádků.
Pak $S_1 \geq S_2$ a rovnost nastává právě když jsou oba kolapsované řádky shodné až na multiplikativní konstantu.

Důsledek: Buď S_1 hodnota statistiky CHISQ v tabulce RxC, S_2 hodnota CHISQ v tabulce rxc (kde $2 \leq r \leq R$, $2 \leq c \leq C$), která vznikla z předchozí kolapsováním. Pak $S_1 \geq S_2$.

Počet stupňů volnosti v kolapsované tabulce je ovšem nižší než v původní, ze srovnání statistik CHISQ proto neplyne nic o síle testů. Nejvyšší dosažená hladina významnosti se může kolapsováním snížit i zvýšit.

Uvažujme nyní RxC tabulku T, ve které je CHISQ≠0, a množinu X všech 2x2 tabulek kolapsovaných z tabulky T. Těchto tabulek je $(2^R-2)(2^C-2)$. V dalším si všimneme vlastností 2x2 kolapsované tabulky s maximální nabývanou hodnotou CHISQ. Nechť t je dále některá tabulka z množiny X, ve které míra PSI /tudíž i CHISQ/ nabývá maximální hodnoty, maximum je uvažováno přes množinu X.

Tvrzení: Nechť v RxC tabulce je prvních V řádků shodných až na multiplikativní konstantu /t.j. jde o týž vektor násobený obecně různými skaláry/, obdobně prvních W sloupců. Nechť CHISQ≠0. Nechť t je kolapsovaná 2x2 tabulka s maximální hodnotou CHISQ. Potom řádky 1,...,V tabulky T byly kolapsovány do téže řádky tabulky t a obdobně sloupce 1,....,W tabulky T byly kolapsovány do téhož sloupce tabulky t.

Shrňme si některé důsledky, které vyplývají z již uvedených tvrzení pro "PSI-optimální" kolapsovanou 2x2 tabulku t:
/a/ Pro libovolnou tabulku $\tau \in X$ platí: t je asociativně lepší než τ nebo t je s τ nesrovnatelná. /srv. odstavec 2.1/.
/b/ Lze-li RxC tabulku T permutacemi řádků a permutacemi sloupců převést na

tvar blokově diagonální matice, pak t má v obou políčkách vedlejší diagonály nuly. /srv.odstavec 2.2/

/c/ Shodné - případně až na multiplikativní konstantu shodné - řádky jsou "kolapsovány společně", obdobně pro sloupce. /srv. poslední tvrzení/.

3. Procedura COLLAPS

3.1. Obecná formulace úlohy

Je dána RxC tabulka (T1) a parametry r,c,N,CRIT,MEASURE; $2 \le r \le R$, $2 \le c \le C$, $N \ge 1$ přirozená, $CRIT \ge 0$ reálné, MEASURE je míra na rxc tabulkách.

Představme si, že všechny možné rxc tabulky kolapsované z tabulky (T1), byly uspořádány do posloupnosti podle klesající hodnoty MEASURE. Z této posloupnosti uvažujeme pouze prvních N tabulek a z nich pak pouze ty s hodnotou MEASURE větší nebo rovnou CRIT.

Uvažované kolapsované tabulky považujeme za řešení úlohy optimálního kolapsování.

První tabulka v tomto seznamu je rxc kolapsovanou tabulkou s největší hodnotou MEASURE. Při rozumné volbě MEASURE /např.CHISQ/ bude v této tabulce - přes zjednodušení původní RxC tabulky - zachováno v jistém smyslu co nejvíce informace o vzájemné závislosti znaků. /srv. odstavec 2.4/. Následující tabulky seznamu pak nabízejí různé alternativní pohledy.

3.2. Nehierarchická procedura

Úlohu formulovanou v předchozím odstavci pro dvourozměrné tabulky by bylo možné snadno formulovat i pro tabulky vícerozměrné, my se však dále budeme věnovat naopak její algoritmizované a implementované restrikci.

Restrikce spóčívá v položení MEASURE=PSI a r=c=2. Pro danou RxC tabulku procedura tady nalezne nejvýše N kolapsovaných 2x2 tabulek optimálních ve smyslu míry PSI.

Příklad. Zkoumejme z literatury známou tabulku /Snee/popisující vztah mezi barvou očí (řádky) a barvou vlasů (sloupce) v severoamerické populaci:

	A- ČERNÝ	B- BRUNET	C- ZRZAVÝ	D- BLONDÝN
1 HNĚDÉ	68	119	26	7
2 ŠEDÉ	15	54	14	10
3 ZELENÉ	5	29	14	16
4 MODRÉ	20	84	17	94

Hodnota statistiky CHISQ je 138.2, tedy významná při devíti stupních volnosti. Nehierarchická procedura COLLAPS zde nabízí následující tři kolapsované tabulky, "nejlepší" z 256 možných:

	D	A,B,C
3,4	110	169
1,2	17	296

PSI = 0.413

	D	A,B,C
4	94	121
1,2,3	33	296

PSI = 0.410

	D	A,B,C
2,3,4	120	252
1	7	213

PSI = 0.342

Hodnota PSI dvou prvních tabulek je přibližně stejná:
obě tabulky vypovídají o souvislosti blond vlasů se světlejšími barvami očí.

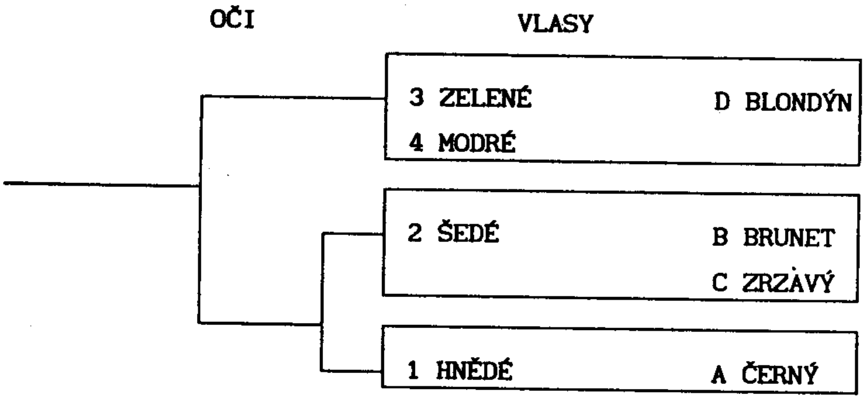
3.3. Hierarchická procedura

První kolapsovaná 2x2 tabulka z řešení uvedeného příkladu hovoří o souvislosti kategorií 3 a 4 s kategorií D na straně jedné a o souvislosti kategorií 1 a 2 s kategoriemi A,B,C na straně druhé. Tázajíce se po podrobnějším popisu souvislosti kategorií 1,2 s kategoriemi A,B,C můžeme provést "rekursivní krok": vytvořit podtabulku 2x3 s řádky 1,2 a sloupci A,B,C a znovu užít nehierarchickou proceduru COLLAPS. Nejlepší kolapsovou 2x2 tabulkou je

	B,C	A
2	68	15
1	145	68

PSI = 0.139

Hierarchická procedura spočívá v rekursivním užití popsané procedury nehierarchické: Podle nejlepší kolapsované tabulky, nalezené v právě zkoumané tabulce (podtabulce) vytvoříme dvě nové podtabulky pro další zkoumání: První podtabulku vytvoříme z kategorií, které přispěly k levému hornímu políčku kolapsované tabulky, druhou podtabulku z kategorií, které přispěly k pravému dolnímu políčku. Pokud se v právě vyšetřované podtabulce nenalezla kolapsovaná tabulka s hodnotou PSI = CRIT, podtabulky se nevytvoří. Podtabulky "degenerované"(1xc nebo rx1) se dále nezkoumají.
Výsledkem hierarchické procedury je korespondence mezi hierarchickými rozklady množin hodnot R={1,...,R} a C={1,...,C}. Tato korespondence je určena nejlepšími kolapsovanými tabulkami nalezenými pro vyšetřované podtabulky. Lze ji graficky vyjádřit buď dendrogramem;



nebo schematem blokového rozkladu tabulky:

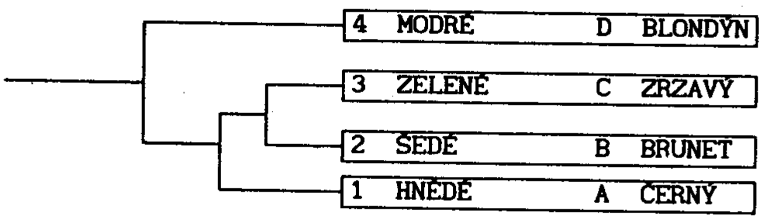
	D	B	C	A
3	2.28	-1.91	6.33	-6.67
4	47.88	-19.86	-8.78	-19.22
2	-9.95	9.08	2.85	-1.96
1	-40.19	12.72	-0.38	27.87

ve kterém můžeme sledovat kladnou souvislost v blocích ve směru hlavní diagonály. Políčka rozkladu mohou být vyplněna buď pozorovanými četnostmi nebo hodnotami některé jiné RxC matice popsané v odstavci 1.3, zde byly užity rozdíly pozorovaných a očekávaných četností $(n_{ij} - e_{ij})$.

3.4. Procedura s orákulem

je hierarchickou procedurou, kde uživatel v každém kroku volí z nabídnutých N kolapsovaných tabulek jednu subjektivně nejvhodnější /"orakulum"/ pro generování podtabulek v dalších krocích . Srv. [15].

V prvním kroku analýzy Sneeovy tabulky byla zvolena druhá kolapsovaná tabulka, celkový rozklad pak byl:



3.5. Konzistence rozkladu

Analyzovaná kontingenční tabulka (n_{ij}) vznikla výběrem z dvourozměrného multinominálního rozdělení s hustotou danou maticí pravděpodobností (p_{ij}).

Na tuto matici můžeme aplikovat popsanou hierarchickou proceduru, otázkou je, zda se optimální kolapsování v teoretické a výběrové struktuře budou shodovat. Obecně se dá ukázat, že pro danou matici pravděpodobností (p_{ij}) a dané $\varepsilon>0$ existuje m_o takové, že pro výběry rozsahu alespoň m_o je optimální kolapsování ve výběrové struktuře také optimálním kolapsováním ve struktuře teoritické /resp. jedním z optimálních kolapsování ve struktuře teoretické/ s pravděpodobností alespoň $1-\varepsilon$.

Zajímavý i užitečný by byl algoritmus, který by se znalostí tabulky pozorovaných četností (n_{ij}) /a neznalostí teoretických četností (p_{ij}) určil "hloubku", do níž máme rozklad v jednotlivých "větvích" provádět tak, aby optimalita takto restrikovaného rozkladu v teoritické struktuře byla zaručena s danou pravděpodobností. Jinými slovy, jde o návrh vhodného zastavovacího pravidla. Návrh nepříliš konzervativního pravidla s popsanou vlastností je otevřenou a patrně netriviální úlohou.

Řadu kroků v tomto směru učinil T.Havránek. Stručně o dvou z nich:

(a) Pro každou generovanou podtabulku se můžeme ptát, zda v ní je evidence závislosti, kterou bychom se pak snažili procedurou COLLAPS vysvětlit. K tomu lze užít běžného testu nezávislosti v dvourozměrné kontingenční tabulce. Takových testů však během hierarchického postupu provedeme více, proto je vhodné užít postupů simultánní inference. Užitečné je zde Holmovo zlepšení Bonferoniovského postupu (podrobněji viz [9]).

(b) K usnadnění posuzování optimality rozkladu sestrojil Havránek následující test: Zvolme pro dvourozměrnou kontingenční tabulku dvě její kolapsování na 2x2 tabulky; míra PSI v těchto tabulkách nabývá hodnot ψ_1 a ψ_2. Označme $\underline{\psi}_1$ a $\underline{\psi}_2$ příslušné hodnoty PSI v teoretické struktuře, tj. hodnoty PSI v 2x2 tabulkách kolapsovaných z tabulky teoretických četností (p_{ij}). Postupem popsaným v [9], testujeme hypotézu $\underline{\psi}_1 = \underline{\psi}_2$ proti alternativě $\underline{\psi}_1 > \underline{\psi}_2$.

Příklad: Srovnáme-li popsaným testem tři tabulky z příkladu 3.2, zjistíme významný rozdíl na hladině 0.05 mezi první a třetí tabulkou a obdobně mezi druhou a třetí tabulkou. Rozdíl mezi první a druhou tabulkou významný není.

3.6. Implementace procedury

Procedura COLLAPS je jednou z metod mechanizované formace hypotéz srv. [8] a je implementována v programovém systému GUHA. Program COLLAPS je napsán v jazyce PL/1 /s překladačem F/. Program COLLAPS - i celý programový systém GUHA - lze implementovat na počítačích řady IBM/370 a počítačích kompatibilních, např. EC1040 nebo EC1033.

Program COLLAPS /viz uživatelskou dokumentaci [16]/ umožňuje užití popsané nehierarchické i hierarchické procedury /srv. odst.3.2 a 3.3/ založené na míře CHI. Proceduru s orákulem /srv. 3.4/ lze simulovat v několika bězích dávkového zpracování. Připravovaná je implementace měr PSI, YATES, W a složitějších zastavovacích pravidel /srv. 3.5 (a),(b)/.

Implementovaný algoritmus pro hledání optimálních kolapsovaných tabulek nezkoumá slepě všechny možnosti /jak činí tzv. "Algoritmy Britského Muzea"/, ale využívá řady triků heuristického charakteru, umožňujících urychlení výpočtu.

Procedura COLLAPS byla s prospěchem užita při analýze fysiologických, lékařských a sociologických dat. Zrychlující techniky se ukázaly dostatečně účinnými; k řešení tabulek běžných rozměrů /cca 10x10/ stačily zlomky minuty na IBM 370/135.

Literatura

[1] J.Anděl: On interactions in contingency tables. Aplikace matematiky 18/1973/, 99-109
[2] J.Anděl: The most significence interaction in a contingency table. Aplikace matematiky 19/1974/, 246-252
[3] M.B.Brown: The indentification of sources of significence in two-way contingency table. Applied Statistics 23/1974/, 405-413
[4] W.J.Dixon, M.B.Brown /Eds./: BMDP-79 Biomedical Computer Programs. University of California Press 1979
[5] K.R.Gabriel: Simultaneous test procedures for multiple comparisons on categorical data. J.Amer.Statist.Assoc. 61/1966/, 1081-1096
[6] L.A.Goodman, W.H.Kruskal: Measures of association for cross-classification. J.Amer.Statist.Assoc. 49/1954/,732-764
[7] S.J.Haberman: The analysis of residuals in cross-classified tables. Biometrics 29/1973/, 205-220
[8] P.Hájek, T.Havránek: Mechanizing hypothesis formation. Mathematical foundation for a general Theory. Springer Verlag /1978/
[9] T.Havránek: Some comments on the GUHA procedures. Explorative datenanalyse, Medizinische Informátik and Statistik vol.26, Springer, Berlin 1981, 156-171
[10] T.Havránek, D.Pokorný: GUHA-style processing of mixed data. Int.J.Man-Machine Studies 9/1977/, 439-447
[11] J.Lancaster: Contingency tables treated by partition κ^2. J.Royal Statist.Soc., Ser.B 5/1971/, 242-259
[12] F.Mosteller: Association and estimation of contingency tables. J.Amer.Statist.Assoc. 63/1968/, 1-28
[13] D.Pokorný: Nominální kalkuly a zobecněné asociační kvantifikátory. Dipl. práce MFF UK, Praha/1975/
[14] D.Pokorný: Non-asymptotical relations between chi-square and interaction tests. Sdělení MSBÚ 12(4). MSBÚ ČSAV,Praha 1976
[15] D.Pokorný: The GUHA method and desk calculators. Intern.J.Man-Machine Studies 10/1978/, 75-86
[16] D.Pokorný: COLLAPS. Uživatelský manuál. MSBÚ Praha/1979/
[17] D.Pokorný: Knowledge acquisition by the GUHA method. Intern.J.Policy Analysis and Information Systems4/1980/, 379-399
[18] D.Pokorný, T.Havránek: On some procedures identifing sources of dependence in contingency tables. COMPSTAT 78, L.C.A.Corsten, J.Herman/Eds/,pp.221-227. Physica-Verlag, Wien/1978/