ZPRACOVÁNÍ KATEGORIÁLNÍCH DAT

Jaroslav Michálek, KAM PřF UJEP
Brno

V příspěvku jsou diskutovány hierarchické, grafové a rozložitelné modely pro popis struktury závislostí mezi faktory pro danou mnohorozměrnou kontingenční tabulku. Pozornost je věnována interpretaci modelů, maximálně věrohodným odhadům pravděpodobnosti vysvětlující model, testování adekvátnosti modelů a výpočetní složitosti. Práce se doplňuje s prací T.Havránka v tomto sborníku.

1. ÚVOD

Nechť $\Gamma = \{\alpha, \beta, \dots, \xi\}$ je množina klasifikačních kritérií nebo faktorů a z nechť je počet jejích prvků. Budeme se zabývat statistickými modely pro analýzu vztahů mezi faktory z množiny Γ.

Zavedeme následující označení: Pro každý faktor $\gamma \in \Gamma$ nechť je $I_\gamma = \{1, 2, \dots, z_\gamma\}$ množina možných hodnot faktoru γ. Jednotlivé hodnoty - úrovně faktoru γ v množině I_γ jsou očíslovány pouze pro jejich identifikaci, předpokládáme, že každý faktor je nominální proměnná. Dále položme $I = \prod_{\gamma \in \Gamma} I_\gamma$ a každý prvek $i = (i_1, \dots, i_z) \in I$ nazveme třídou. Počet tříd označíme K a tedy $K = \prod_{\gamma \in \Gamma} z_\gamma$.

Nechť X_γ, je hodnota faktoru γ, $\gamma \in \Gamma$ na náhodně vybraném objektu. Budeme předpokládat, že náhodný vektor $X = (X_\alpha, X_\beta, \dots, X_\xi)$ má diskrétní rozdělení pravděpodobností (I, p) s oborem hodnot I a pravděpodobnostní funkcí $p(i) = P(X_\alpha = i_1, \dots, X_\xi = i_z)$ $i \in I$ a nechť $p(i) > 0$ pro každé $i \in I$.

Předpokládejme, že nezávisle klasifikujeme n objektů, tj. budeme předpokládat, že je dán náhodný výběr rozsahu n z rozdělení (I, p). Potom pro každou třídu $i \in I$ můžeme stanovit její četnost $N(i)$, jako počet prvků ve výběru, které byly při klasifikaci podle faktorů z Γ zařazeny do třídy i.

Vektor $(N(i), i \in I)$ má zřejmě multinomické rozdělení pravděpodobností

$$P(N(i) = n(i), i \in I) = \frac{n!}{\prod_{i \in I} n(i)!} \prod_{i \in I} p(i)^{n(i)} \quad , \quad n(i) \geq 0 \; , \; \sum_{i \in I} n(i) = n \; ,$$

kde pravděpodobnosti $p(i)$, $i \in I$ jsou neznámé parametry.

Dále budeme veličiny $N(i)$ a jejich realizace značit $n(i)$, z kontextu bude jasné, zda se jedná o náhodnou veličinu, či její realizaci.

Pomocí četností $n(i)$ zavedeme z-rozměrnou kontingenční tabulku jako množinu $\{n(i), i \in I\}$.

Klasifikací objektů pouze podle faktorů z množiny $a \subseteq \Gamma$ dospějeme k zavedení marginální třídy $i_a \in I_a = \prod_{\gamma \in a} I_\gamma$ a odpovídající marginální četnosti $n(i_a) = \sum_{j: j_a = i_a} n(j)$ třídy i_a a dále k marginální kontingenční tabulce $\{n(i_a), i_a \in I_a\}$.

Při analýze vztahů mezi faktory z množiny Γ se budeme především zajímat o jejich nezávislost a podmíněnou nezávislost. Pojem nezávislosti je zřejmý. Podmíněnou nezávislost faktorů $\alpha \in \Gamma$ a $\beta \in \Gamma$ při dané hodnotě faktoru $\gamma \in \Gamma$ definujeme vztahy $P(X_\alpha = i_\alpha, X_\beta = i_\beta \mid X_\gamma = i_\gamma) = P(X_\alpha = i_\alpha \mid X_\gamma = i_\gamma) \cdot P(X_\beta = i_\beta \mid X_\gamma = i_\gamma)$ pro všechna $i_\alpha \in I_\alpha$, $i_\beta \in I_\beta$ a $i_\gamma \in I_\gamma$ taková, že $P(X_\gamma = i_\gamma) > 0$. Budeme ji zkráceně označovat $\alpha \perp \beta \mid \gamma$.

Analogicky definujeme podmíněnou nezávislost faktorů z množiny $a \subseteq \Gamma$ a faktorů z množiny $b \subseteq \Gamma$ při dané množině faktorů $c \subseteq \Gamma$ a označujeme ji $a \perp b \mid c$.

Vlastnosti podmíněné nezávislosti později využijeme při názorné interpretaci tzv. grafových modelů pro neznámou pravděpodobnost p.

2. LOGARITMICKO-LINEÁRNÍ MODELY PRO p(i)

Dále popsaný přístup ke konstrukci zaváděných modelů je založen na teorii tzv. log-lineárních interakčních modelů. Viz [7],[3],[4], [0].

Předpokládejme, že $\mathcal{S}$ je neprázdný systém podmnožin množiny Γ, tj. $\emptyset \neq \mathcal{S} \subseteq 2^{\Gamma}$. Budeme říkat, že pravděpodobnost p se řídí log-lineárním interakčním modelem $\mathcal{M}_{\mathcal{S}}$ a psát $p \in \mathcal{M}_{\mathcal{S}}$, existují-li funkce $\lambda_a(i_a)$, $a \in \mathcal{S}$ takové, že platí

$$\log p(i) = \sum_{a \in \mathcal{S}} \lambda_a(i_a) \tag{1}$$

pro každé $i \in I$. Funkce λ_a se potom nazývají efekty nebo též interakce, klademe $\lambda_{\emptyset}$ = = konst. Model $\mathcal{M}_{\mathcal{S}}$ daný vzorcem (1) se někdy nazývá též obecný log-lineární model, protože zde nejsou kladeny žádné požadavky na volbu systému $\mathcal{S}$. Jeho relativní výhodou snad je, že formálně lze snadno přecházet k tabulkám velké dimenze, ale velkou nevýhodou potom je jeho interpretace. Už pro dimenze $z \geq 4$ je těžké zinterpretovat, co daný model vlastně znamená. Také formální zápis tohoto obecného modelu je poněkud těžkopádný. Proto se obvykle přechází ke třídě jednodušších modelů, kterou jsou hierarchické log-lineární interakční modely, stručně budeme říkat jenom hierarchické modely. Tyto modely jsou jednoznačně určeny jejich generující sentencí $\mathcal{A}$.

2.1. Hierarchické modely

Budeme říkat, že neprázdný systém $\mathcal{A} = \{a_1, \ldots, a_k\}$ podmnožin množiny Γ je generující sentencí, když v $\mathcal{A}$ neexistuje prvek a, který by byl podmnožinou nějakého jiného prvku z $\mathcal{A}$. Dále budeme říkat, že $\mathcal{M}_{\mathcal{A}}$ je hierarchický model s generující sentencí $\mathcal{A}$ pro pravděpodobnost p a psát $p \in \mathcal{M}_{\mathcal{A}}$, jestliže platí

$$\log p(i) = \sum_{a \subseteq \Gamma} \lambda_a(i_a) \quad , \text{ pro každé } i \in I ,$$

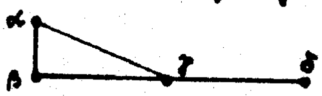
pro nějaké funkce λ_a takové, že $\lambda_a \equiv 0$, když neexistuje $c \in \mathcal{A}$ takové, že $a \subseteq c$.

Generující sentence $\mathcal{A}$ hierarchického modelu $\mathcal{M}_{\mathcal{A}}$ umožňuje jednoduchý zápis modelu a dobře se s ní pracuje. Dále je možné přiřadit každé generující sentenci $\mathcal{A}$ (a tedy také každému hierarchickému modelu $\mathcal{M}_{\mathcal{A}}$) jednoduchý neorientovaný graf $G_{\mathcal{A}} = (V,H)$, kde množina uzlů $V = \bigcup_{a \in \mathcal{A}} a$ a množina hran $H = \{(\alpha,\beta) : \exists \ a \in \mathcal{A} : \{\alpha, \beta\} \subseteq a\}$. Uvedenou situaci budeme ilustrovat příkladem:

Nechť z=4, $\Gamma = \{\alpha, \beta, \gamma, \delta\}$ a uvažujme generující sentenci $\mathcal{A} = \{\{\alpha,\beta,\gamma\},\{\gamma,\delta\}\}$, kterou pro jednoduchost zapíšeme zkráceně ve tvaru $\mathcal{A} = \{\alpha\beta\gamma, \gamma\delta\}$ (a tohoto zkráceného způsobu zápisu budeme využívat i dále). Pak pravděpodobnost $p \in \mathcal{M}_{\mathcal{A}}$ je tvaru

$$\log p(i) = \lambda_{\emptyset} + \lambda_{\alpha}(i_{\alpha}) + \lambda_{\beta}(i_{\beta}) + \lambda_{\gamma}(i_{\gamma}) + \lambda_{\alpha\beta}(i_{\alpha\beta}) + \lambda_{\beta\gamma}(i_{\beta\gamma}) + \lambda_{\alpha\gamma}(i_{\alpha\gamma}) + \lambda_{\alpha\beta\gamma}(i_{\alpha\beta\gamma}) + \lambda_{\delta}(i_{\delta}) + \lambda_{\gamma\delta}(i_{\gamma\delta}). \tag{2}$$

Odpovídající graf $G_{\mathcal{A}} = (V,H)$ je tvaru :



Snadno nahlédneme, že dva různé hierarchické modely $\mathcal{M}_{\mathcal{A}}$, $\mathcal{M}_{\mathcal{A}'}$ mohou mít stejný graf. Kdybychom v předchozím příkladě zavedli místo generující sentence $\mathcal{A}$ generující sentenci $\mathcal{A}' = \{\alpha\beta, \alpha\gamma, \beta\gamma, \gamma\delta\}$, ihned bychom mohli zjistit, že $\mathcal{M}_{\mathcal{A}} \neq \mathcal{M}_{\mathcal{A}'}$, ale $G_{\mathcal{A}} = G_{\mathcal{A}'}$. Tedy daný hierarcický model není jednoznačně určen svým grafem. Tato skutečnost působí těžkosti při interpretaci hierarchického modelu pouze na základě jeho grafu. Proto se zavádí užší třída hierarchických modelů taková, že každý její model je jednoznačně reprezentován svým grafem.

2.2. Grafové modely

Nejdříve zavedeme pojem markovské pravděpodobnosti vzhledem k danému jednoduchému neorientovanému grafu $G = (V,H)$, $V \subseteq \Gamma$. Tento pojem nám umožní zavést třídu grafových modelů a zároveň jejich jednoduchou a názornou interpretaci, kterou lze vyčíst jenom z grafu příslušného modelu.

Budeme říkat, že pravděpodobnost p je markovská vzhledem ke grafu $G = (V,H)$, jestliže platí

1. $p(i) > 0$ pro každé $i \in I$
2. $p(i_\gamma) = 1/z_\gamma$ pro každé $\gamma \notin V$
3. Pro libovolné dva uzly $\alpha, \beta \in V$, které nejsou sousední (tj. nejsou spojeny hranou) platí $\alpha \perp \beta \mid V - \{\alpha, \beta\}$.

Dále říkáme, že p je rozšířená markovská pravděpodobnost, když $p(i) = \lim_{m \to \infty} p_m(i)$, $i \in I$ a pravděpodobnosti p_m jsou markovské.

Z této definice je zřejmé, že rozšířená markovská pravděpodobnost si zachovává obě vlastnosti 2. a 3. z definice markovské pravděpodobnosti, které jsou zvlášť důležité pro interpretaci. Další vlastnosti markovské pravděpodobnosti, jež se dají při interpretaci daného grafu dobře využít, jsou dány následující větou (viz [11]).

VĚTA 1: Následujících 5 tvrzení o pravděpodobnosti p je ekvivalentních:

1. p je markovská vzhledem ke grafu $G = (\Gamma, H)$
2. Existují funkce λ_a, $a \subseteq \Gamma$ takové, že $\log p(i) = \sum_{a \subseteq \Gamma} \lambda_a(i_a)$, $i \in I$ a $\lambda_a(i_a) \equiv 0$ jen když a není úplná podmnožina vrcholů grafu G. (Množina $a \subseteq \Gamma$ se nazývá úplná, když každé dva její vrcholy jsou spojeny hranou).
3. Když množina vrcholů $a \subseteq \Gamma$ odděluje uzly α a $\beta \in \Gamma$ (tj. když každá cesta z α do β vede přes a), pak $\alpha \perp \beta \mid a$.
4. Pro každé $a \subseteq \Gamma$ platí $a \perp a' \mid \partial a$, kde $a' = \Gamma - a$ a ∂a značí hranici a tj.: $\partial a = \{\beta \in \Gamma - a : \exists\ \alpha \in a : \alpha \sim \beta\}$, zde $\alpha \sim \beta$ značí, že uzly α a β jsou v G sousední.
5. Pro každé $\alpha \in \Gamma$ platí $\alpha \perp \alpha' \mid \partial\alpha$, kde místo $\{\alpha\}$ píšeme pouze α.

PŘÍKLAD: Uvažujme hierarchické modely $\mathfrak{M}_{\mathcal{A}}$ a $\mathfrak{M}_{\mathcal{A}'}$, které odpovídají generujícím sentencím $\mathcal{A}$ a $\mathcal{A}'$ uvažovaným v předchozím příkladě. Jim odpovídají pravděpodobnosti p (je daná vzorcem (2)) a p' tvaru

$$\log p'(i) = \lambda'_\emptyset + \lambda'_\alpha(i_\alpha) + \lambda'_\beta(i_\beta) + \lambda'_\gamma(i_\gamma) + \lambda'_\delta(i_\delta) + \lambda'_{\alpha\beta}(i_{\alpha\beta}) + \lambda'_{\alpha\gamma}(i_{\alpha\gamma}) + \lambda'_{\beta\gamma}(i_{\beta\gamma}) + \lambda'_{\delta\gamma}(i_{\delta\gamma}), \qquad i \in I.$$

Snadno nahlédneme, že pravděpodobnost p splňuje podmínku 2 z věty 1 a je tedy markovská, zatímco pravděpodobnost p' tuto podmínku nesplňuje neboť množina $\{\alpha\beta\gamma\}$ je úplná v grafu $G_{\mathcal{A}'}$ a tedy $\lambda'_{\alpha\beta\gamma} \equiv 0$, ale pravděpodobnost p' není markovská vzhledem ke grafu $G_{\mathcal{A}'}$. Tedy interpretaci modelu $\mathfrak{M}_{\mathcal{A}'}$ není možné provést pouze z grafu $G_{\mathcal{A}'}$ pomocí vlastností markovské pravděpodobnosti a tedy jen pomocí podmíněné nezávislosti na základě ekvivalencí 1<=>3<=>4<=>5 z věty 1. Na druhé straně lze uvedených ekvivalencí s výhodou použít při interpretaci modelu $\mathfrak{M}_{\mathcal{A}}$. Tak např. z grafu $G_{\mathcal{A}}$ pomocí vlastnosti 4 ihned vidíme, že faktory α a β jsou podmíněně nezávislé s faktorem δ při daném faktoru γ tj. $\{\alpha, \beta\} \perp \delta \mid \gamma$.

Uvedený příklad motivuje zavedení grafových modelů: Řekneme, že hierarchický model $\mathfrak{M}_{\mathcal{A}}$ je grafový, jestliže pravděpodobnost p je markovská vzhledem ke grafu $G_{\mathcal{A}}$.

O tom, zda daný model $\mathfrak{M}_{\mathcal{A}}$ je grafový, lze rozhodnout na základě věty (viz [11], [3]).

VĚTA 2: Model $\mathfrak{M}_{\mathcal{A}}$ je grafový právě když $\mathcal{A}$ je množina všech klik grafu $G_{\mathcal{A}}$. (Klika grafu je úplná a vzhledem k inkluzi maximální podmnožina uzlů).

Z uvedené věty je ihned patrné, že model $\mathfrak{M}_{\mathcal{A}}$ uvedený výše je grafový, neboť $\mathcal{A}$ obsahuje obě kliky $\{\alpha, \beta, \gamma\}$ a $\{\gamma, \delta\}$ grafu $G_{\mathcal{A}}$, zatímco model $\mathfrak{M}_{\mathcal{A}'}$ není grafový, neboť $\mathcal{A}'$ neobsahuje kliku $\{\alpha, \beta, \gamma\}$ grafu $G_{\mathcal{A}'}$.

Jednoduchá a názorná interpretace grafového modelu $\mathfrak{M}_{\mathcal{A}}$ pomocí jeho grafu $G_{\mathcal{A}}$ vede k tomu, že v poslední době se při vyhledávání vhodného modelu pro vysvětlení dané kontingenční tabulky (tj. pro popis pravděpodobnosti p) stále více používá třídy grafových modelů. Při tom hraje ještě důležitou roli okolnost, že třída grafových modelů je uzavřená vzhledem ke konjunkci modelů (viz [7]). Částečnou nevýhodou třídy grafových modelů je skutečnost, že odhad neznámé pravděpodobnosti p v grafovém modelu není možné vždy stanovit přímo, pomocí příslušných marginálních četností, ale je nutné je hledat speciálním iteračním algoritmem (viz odstavec 3). Dále popíšeme podtřídu grafových modelů tzv. třídu rozložitelných modelů, v níž je možné stanovit odhad neznámé pravděpodobnosti p bez pomoci zmíněného iteračního algoritmu.

2.3. Rozložitelné modely

Pojem rozložitelného modelu pochází od Habermana [6]. Řekneme, že hierarchický model $\mathcal{M}_{\mathcal{C}}$ je rozložitelný, jestliže existují generující sentence $\mathcal{A}$ a $\mathcal{B}$ tak, že $\mathcal{A} \cup \mathcal{B} = \mathcal{C}$, $\mathcal{A} \cap \mathcal{B} = \emptyset$ a $(\bigcup_{a \in \mathcal{A}} a) \cap (\bigcup_{b \in \mathcal{B}} b) = a^* \cap b^*$ pro nějaké $a^* \in \mathcal{A}$ a $b^* \in \mathcal{B}$.

Lze ukázat, že každý rozložitelný model je grafový (viz [12]) a dále (viz [3]) je možné zavést ekvivalentní definici rozložitelného modelu následujícím způsobem:

$\mathcal{M}_{\mathcal{C}}$ je rozložitelný právě když existuje takové uspořádání prvků jeho generující sentence $\mathcal{C} = \{a_1, \dots, a_k\}$, že $a_t \cap \{a_1 \cup \dots \cup a_{t-1}\} = a_t \cap a_{r_t}$ a $r_t \in \{1, \dots, t-1\}$ pro $t = 2, 3, \dots, k$.

Užijeme-li této definice rozložitelnosti, lze snadno odvodit (viz [3]), že pravděpodobnost p je tvaru

$$p(i) = \frac{\prod_{t=1}^{k} p(i_{a_t})}{\prod_{t=2}^{k} p(i_{b_t})}, \quad i \in I \tag{3}$$

kde $b_t = a_t - a_{r_t}$, $t = 2, \dots, k$ a $p(i_a)$ značí marginální pravděpodobnost, tj.

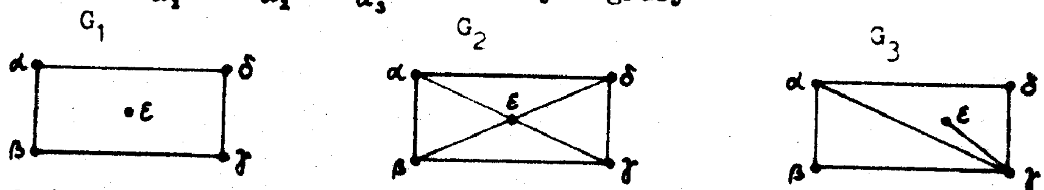
$$p(i_a) = \sum_{j: j_a = i_a} p(j), \quad a \subset \Gamma.$$

Multiplikativního rozkladu (3) pravděpodobnosti p lze využít ke stanovení odhadu pravděpodobnosti p (viz odstavec 3).

Protože rozložitelný model je grafový, vzniká otázka, jak z daného grafu poznáme, že odpovídající model $\mathcal{M}_{\mathcal{A}}$ je rozložitelný. Odpověď je dána následující větou (viz [3])

VĚTA 3: Model $\mathcal{M}_{\mathcal{A}}$ je rozložitelný právě když graf jeho generující sentence je triangulový (tj. neobsahuje cykly bez úhlopříček délky čtyři nebo větší).

Větu 3 budeme ilustrovat příkladem.: Uvažujme množinu faktorů $\Gamma = \{\alpha, \beta, \gamma, \delta, \varepsilon\}$ a grafové modely $\mathcal{M}_{\mathcal{A}_1}$, $\mathcal{M}_{\mathcal{A}_2}$, $\mathcal{M}_{\mathcal{A}_3}$ zadané svými grafy



Pak z věty 3 je ihned patrné, že modely s grafy G_1 a G_2 nejsou rozložitelné (obsahují 4-cykl $\alpha \sim \beta \sim \gamma \sim \delta \sim \alpha$ bez úhlopříček), ale model s grafem G_3 rozložitelný je.

3. MAXIMÁLNĚ VĚROHODNÝ ODHAD PRAVDĚPODOBNOSTI p

Obecný algoritmus tzv IPF (Iterative Proportional Fitting) pro nalezení maximálně věrohodného odhadu neznámé pravděpodobnosti p v hierarchickém modelu $\mathcal{M}_{\mathcal{A}}$ je popsán v [2]. Zde se zaměříme na odhad p v grafových modelech. Algoritmus, který uvedeme, je speciálním případem IPF algoritmu pro situaci, že $\mathcal{A}$ je generující sentencí grafového modelu. Lze ukázat (viz [11]), že maximálně věrohodný odhad $\hat{p}$ pravděpodobnosti p v grafovém modelu $\mathcal{M}_{\mathcal{A}}$ je jediná rozšířená markovská pravděpodobnost $\hat{p}$, která vyhovuje systému rovnic $\hat{p}(i_a) = n(i_a)/n$, $i_a \in I_a$, $a \in \mathcal{A}$.

Pro výpočet p nejsou uvedené rovnice vhodné a užívá se následující algoritmus (viz [11]): Zavedeme operátor

$$(T_a p)(i) = p(i) \frac{n(i)}{n p(i_a)}, \quad i \in I.$$

Dále pro generující sentenci grafového modelu $\mathcal{A} = \{a_1, \dots, a_k\}$ zavedeme operátor $S = T_{a_k} \cdot \ldots \cdot T_{a_1}$ a pomocí něho pravděpodobnost $p_m = S^m p_0$, kde $p_0(i) = 1/K$, $i \in I$ je počáteční aproximace a S^m je m-tá mocnina operátoru S. Pak platí (viz [11]), že

$$\hat{p} = \lim_{m \to \infty} p_m. \tag{4}$$

Při praktickém výpočtu předepíšeme kladné ε (např. $\varepsilon = 0{,}01$ apod.) a $\hat{p}$ aproximujeme hodnotou p_m takovou, že $|p_{m-1}(i) - p_m(i)| < \varepsilon$ pro každé $i \in I$.

Je-li model $\mathcal{M}_{\mathcal{a}}$ rozložitelný, lze - jak bylo poznamenáno dříve - odhadnout pravděpodobnost p přímo pomocí marginálních četností. Lze postupovat dvěma způsoby:

1. První postup je založen na vzorci (3), využívá se relativních marginálních četností $n(i_{a_t})/n$ a $n(i_{b_t})/n$. Po jejich dosazení do (3) za odpovídající marginální pravděpodobnosti $p(i_{a_t})$ a $p(i_{b_t})$ dostaneme příslušný maximálně věrohodný odhad $\hat{p}$. Při tomto postupu odhadu $\hat{p}$ je tedy v obecném případě potřeba testovat rozložitelnost modelu (viz procedura v[7]) a potom nalézt uspořádání prvků generující sentence $\mathcal{a}$, které umožňuje rozklad (3). Výpočetní složitost těchto procedur je pro systém programů, o němž bude řeč v závěru, polynomiální s vedoucími členeny k^3z a $3k^2z$ (viz [17]) a ukazuje se, že je časově výhodnější užít pro odhad p v grafových modelech test rozložitelnosti a potom vzorec (3), než vycházet u všech grafových modelů z algoritmu IPF. Efektivnost tohoto postupu se zvlášť projeví při vyhledávání vhodného grafového modelu nějakou automatickou procedurou (viz např. procedurou uvedenou v [8]).

2. Druhý postup maximálně věrohodného odhadu p v rozložitelném modelu je založen na tzv. indexu $\nu(c)$ souvislého grafu vzhledem k jeho dané úplné množině uzlů c. Index $\nu(c)$ byl zaveden v práci [12] rovněž je definován v [3]. Pomocí něho lze maximálně věrohodný odhad p v modelu $\mathcal{M}_{\mathcal{a}}$ zapsat v explicitním tvaru

$$\hat{p} = \prod_{t=1}^{T} \prod_{c \in \mathcal{a}_t} n(i_c)^{\nu_t(c)}/n^T \quad , \tag{5}$$

kde $\mathcal{a}_t$, $t = 1,\dots,T$ jsou generující sentence souvislých komponent G_t grafu $G_{\mathcal{a}}$ a $\nu_t(c)$ je index souvislé komponenty G_t vzhledem k $c \in \mathcal{a}_t$.

Užití vztahu (5) pro odhad p je podmíněno nalezením rychlého algoritmu pro výpočet $\nu_t(c)$. Naše dosavadní zkušenosti zatím ukazují, že výpočet založený na vzorci (3) je rychlejší.

4. TESTOVÁNÍ ADEKVÁTNOSTI MODELU

V tomto příspěvku si všimneme dvou přístupů k testování adekvátnosti modelu. První a snad nejčastější vychází z klasického testu poměrem věrohodnosti, druhý je založen na teorii zobecněných log-lineárních interakcí. Existují i další přístupy, např. pomocí přesných testů se značnou výpočetní složitostí - viz např. [5] a příspěvek M. Hartmanna v tomto sborníku nebo přístupy jež vycházejí z práce [13].

4.1. Test poměrem věrohodnosti

Uvažujme dva hierarchické modely $\mathcal{M}_{\mathcal{a}}$ a $\mathcal{M}_{\mathcal{a}_0}$ pro pravděpodobnost p (nemusí být grafové). A nechť platí $\mathcal{M}_{\mathcal{a}_0} \subseteq \mathcal{M}_{\mathcal{a}}$ (tj. každý prvek generující sentence $\mathcal{a}_0$ je podmnožinou nějakého prvku generující sentence $\mathcal{a}$). Cílem je testovat hypotézu $p \in \mathcal{M}_{\mathcal{a}_0}$ za podmínky, že $p \in \mathcal{M}_{\mathcal{a}}$. Nemáme-li žádnou informaci o pravděpodobnosti p, obvykle volíme za model $\mathcal{M}_{\mathcal{a}}$ model saturovaný, tj. model s generující sentencí $\mathcal{a} = \{\Gamma\}$.

Test poměrem věrohodnosti uvedené hypotézy je potom založen na dobře známé statistice $h^2 = 2\sum_{i \in I} n(i)\hat{p}(i)/\hat{p}_0(i)$, kde $\hat{p}$ a $\hat{p}_0$ jsou maximálně věrohodné odhady p za předpokladu, že $p \in \mathcal{M}_{\mathcal{a}}$ a $p \in \mathcal{M}_{\mathcal{a}_0}$.

Je dobře známo, že statistika h^2 je asymptoticky ekvivalentní se statistikou

$$\chi^2 = \sum_{i \in I} (n\hat{p}(i) - n\hat{p}_0(i))^2/n\hat{p}_0(i) \; .$$

Obě statistiky mají asymptoticky χ^2 rozdělení s počtem stupňů volnosti $\nu = d_{\mathcal{a}} - d_{\mathcal{a}_0}$, kde

$$d_{\mathcal{a}} = \sum_{t=1}^{k} K_{a_t} - \sum_{t<t'}\sum K_{a_t \cap a_{t'}} + \dots + (-1)^{k-1} K_{a_1 \cap \dots \cap a_k} \tag{6}$$

pro $\mathcal{a} = \{a_1,\dots,a_k\}$ a $K_a = \prod_{\gamma \in a} z_\gamma$, $K_\emptyset = 1$. Pro saturovaný model je $d_{\mathcal{a}} = K$.

Otázkou zůstává, kterou ze statistik h^2 a χ^2 použít, když jsou četnosti n(i) „malé". Této problematice je věnována řada prací. V [10] je teoreticky zdůvodněno, proč je použití statistiky χ^2 při malých četnostech nevhodné. Rudás v [16] provedl simulaci kritických hodnot a intervalů spolehlivosti pro kritické hodnoty správného rozdělení statistik h^2 a χ^2 při malých četnostech $n(i)$ v tabulkách typu 2x2x2, 3x2x2 apod. Ukázalo se, že 90 až 95% těchto intervalů pokrylo 90% a 95% kritické hodnoty Pearsonova χ^2 rozdělení pro velice malé rozsahy výběrů. I když provedená simulace vyznívá optimisticky pro použití statistiky χ^2, bude v praxi při malých četnostech potřeba velké obezřetnosti při jejím použití a spíše vycházet z práce [19].

4.2. Testování pomocí zobecněných interakcí

Uvažujme daný hierarchický model $\mathcal{M}_a$ s generující sentencí $a = \{a_1, \dots, a_k\}$. a-zobecněnou log-lineární interakci $\delta_a(i)$, $i \in I$ zavedeme vztahem

$$\delta_a(i) = \log p(i) - \sum_{t=1}^{k} \log p(i_{a_t}, z_{a_t'}) + \sum_{t<t'}\sum \log p(i_{a_t \cap a_{t'}}, z_{(a_t \cap a_{t'})'}) -$$

$$- \dots + (-1)^k \log p(i_{a_1 \cap \dots \cap a_k}, z_{(a_1 \cap \dots \cap a_k)'}), \qquad (7)$$

kde (i_a, z_a) značí třídu $j = (j_1, \dots, j_z) \in I$, pro kterou platí $j_\gamma = i_\gamma$ pro $\gamma \in a$
$j_\gamma = z_\gamma$ pro $\gamma \notin a$.

Pomocí výsledků práce [4] lze ukázat, že $p \in \mathcal{M}_a$ právě když $\delta_a(i) = 0$ pro každé $i \in I$. Z této této myšlenky vychází test hypotézy $p \in \mathcal{M}_a$ pomocí zobecněných log-lineárních interakcí.

Především lze odvodit, že mezi zobecněnými (log-lineárními) interakcemi $\delta_a(i)$ danými vztahem (7) je právě d_a interakcí, které jsou identicky rovny nule a pro zbylých $\nu_a = K - d_a$ je třeba provést test simultánní hypotézy $\delta_a(i) = 0$. Test, který uvedeme dále je podrobně popsán v [14] a vychází z prací [1], [9].

Označme nejdříve zobecněné interakce, které mají být testovány $\delta_1, \dots, \delta_{\nu_a}$. Z vyjádření (7) plyne, že je lze zapsat ve tvaru $\delta_t = \sum_{j \in I} \varphi_t(j) \log p(j)$, $t = 1, \dots, \nu_a$, kde $\varphi_t(j)$, $j \in I$, $t = 1, \dots, \nu_a$ jsou konstanty dané vyjádřením (7) a $\sum_{j \in I} \varphi_t(j) = 0$.

Označíme-li dále $d_t = \sum_{j \in I} \varphi_t(j) \log n(j)$ výběrové protějšky δ_t, $\alpha \in (0,1)$

$s_t^2 = \sum_{j \in I} \varphi_t^2(j) n / n(j)$ a $c_\alpha = u(0.5 + 0.5(1 - \alpha)^{1/\nu_a})$, kde u je kvantilová funkce normálního rozdělení $N(0,1)$, pak lze ukázat (viz [14]), že intervaly

$$(d_t - c_\alpha s_t, \; d_t + c_\alpha s_t) \qquad (8)$$

jsou $100(1 - \alpha)$% simultánní intervaly spolehlivosti pro δ_t, $t = 1, \dots, \nu_a$. Hypotézu $p \in \mathcal{M}_a$ pak zamítneme na hladině významnosti α, když existuje alespoň jeden interval mezi intervaly (8), který nezahrne nulu.

5. PŘÍKLADY

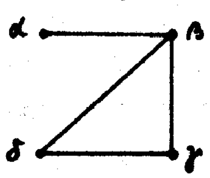
Dále uvedené příklady byly počítány pomocí systému programů, který vytvořila M. Řebíčková (viz [17]), jejich stručný popis je v [15]. V současné době je tento systém rozšířen o možnost interaktivní komunikace.

PŘÍKLAD 1: Porovnání testovacích procedur z odstavců 4.1 a 4.2 bylo provedeno pro kontingenční tabulku TAB 1, původně publikovanou v [18]. Whittaker v [20] analyzoval tuto tabulku metodou založenou na aditivních odchylkách (viz též [21]) a dospěl k výsledku, že tabulku lze vysvětlit grafovými modely $\mathcal{M}_{a_1}$, $\mathcal{M}_{a_2}$, $\mathcal{M}_{a_3}$ uvedenými dále. Tyto modely byly zpracovány metodami odstavce 4.1 a 4.2 a byly získány následující výsledky :

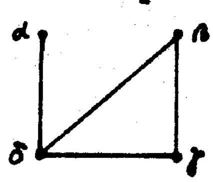
TAB 1

		γ	+	+	−	−
α	β	δ	+	−	+	−
+	+		42	23	6	25
+	−		6	24	7	38
−	+		1	4	1	6
−	−		2	9	2	20

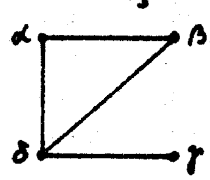
$\mathcal{M}_{a_1}$



$\mathcal{M}_{a_2}$

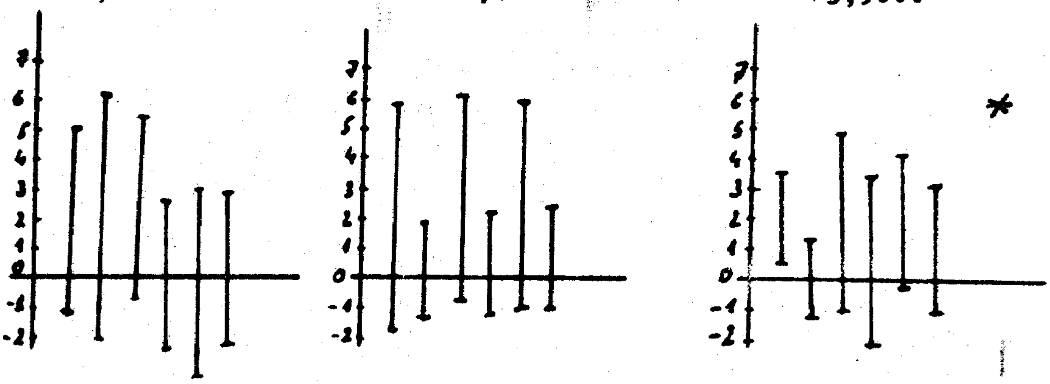


$\mathcal{M}_{a_3}$



Model	$\mathcal{M}_{a_1}$	$\mathcal{M}_{a_2}$	$\mathcal{M}_{a_3}$
Stupně volnosti	6	6	6
χ^2	6,9400	11,6725	14,3518*
h^2	8,0099	11,4281	13,5000

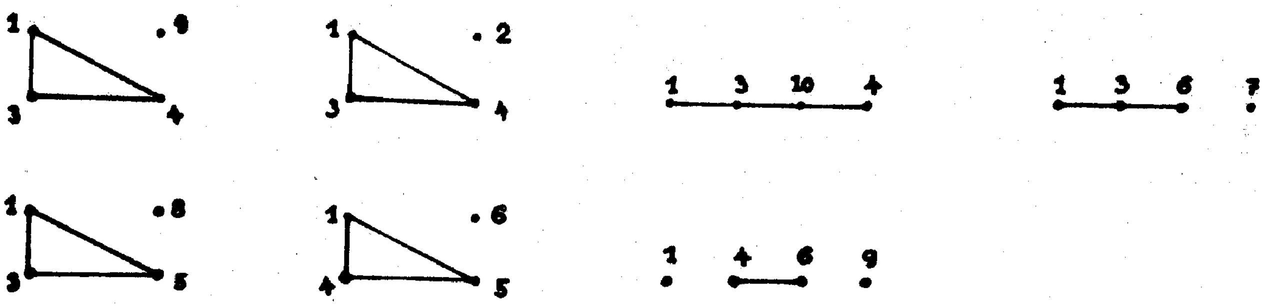
Metoda zobecněných interakcí



V uvedeném přehledu výsledků značí * významnost na hladine významnosti 5%. Z přehledu je vidět, že test založený na zobecněných log-lineárních interakcích se choval stejně jako χ^2 test.

Dále byla pro data z TAB 1 užita strategie pro výběr optimálních adekvátních modelů, která byla navržena Havránkem v [8] . Tato metoda vybrala při použití statistiky χ^2 model $\mathcal{M}_{a_1}$ a při použití statistiky h^2 a také při použití testu založeného na zobecněných log-lineárních interakcích modely $\mathcal{M}_{a_1}$ a $\mathcal{M}_{a_2}$.

PŘÍKLAD 2 je převzatý z práce[17], jeho cílem je demonstrovat praktické užití popisované problematiky. Data pocházejí ze zdravotnicko-psychologického výzkumu. Na 127 respondentech bylo měřeno 9 psychologických faktorů : 2- emotivita, 3-primarita, 4-aktivita, 5- vědomí široké, 6-maskulinita, 7-avidita, 8-smyslové zaměření, 9-citovost a 10-rozumové zaměření a dále faktor 1-zdravotní stav. Účelem bylo dát prvotní informaci o vlivu uvedených psychologických faktorů na zdravotní stav. Vzhledem k malému rozsahu výběru bylo po předběžné analýze rozhodnuto vycházet ze čtyřrozměrných kontingenčních tabulek, které budou vždy zahrnovat faktor 1. Pro tyto tabulky byly procedurou navrženou v [8] hledány adekvátní vysvětlující modely. Dále jsou v grafické formě uvedeny získané výsledky. Jsou z nich dobře patrné interpretační možnosti vyložené problematiky. Na hladině významnosti α = 0,05 byly vybrány tyto modely:



LITERATURA

[1] Anděl J.(1973) On interactions in contingency tables, Aplikace matematiky 18,99-109

[2] Bishop Y.M.M., Fienberg S.E., Holland P.W.(1975) Discrete multivariate analysis: Theory and practice, MIT Press, Cambridge, Mass.

[3] Daroch J., Lauritzen S., Speed T. (1980) Markov fields and log-linear interaction models for contingency tables. Annals of Stat. 8,522-539

[4] Daroch J., Speed T.(1979) Multiplicative and additive models and interactions. Research Report no 49 , Dept. of Theoret. Stat., Arhus

[5] Edwards D.(1984) A computer intensive approach to the analysis of sparse multidimensional contingency tables. COMPSTAT 1984 Proceedings in Computional Statistics - 6-th Symposium, Prague

[6] Haberman S.J.(1974)The analysis of frequency data. IMS monographs,Univ. of Chicago Press
[7] Havránek T.(1982)O analýze mnohorozměrných kontingenčních tabulek. ROBUST 82, Sborník prací ZŠ JČSMF ,pražské pobočky, Podkost
[8] Havránek T. (1984) A Procedure for Model Search in Multidimensional Contingency Tables. Biometrics to appear
[9] Havránek T.(1978) On simultaneous inference in multidimensional contingency tables. Aplikace matematiky 23,31-38
[10] Havránek T.,Edwards D.(1984)A fast procedure for model search in multidimensional contingency tables. RECKU Dokum.,Rapport 84/1,København
[11] Lauritzen S. L.(1982) Lectures on contingency tables. Alborg Univ. Press
[12] Lauritzen S.L.,Speed T.,Vijayan K.(1978) Decomposable graphs and hypergraphs. UCIMS,Preprint No.9
[13] Margolin L. (1974) JASA,69, 755-764
[14] Michálek J.(1984) Goodness of fit tests in multidimensional contingency tables. Colloquium on goodness of fit,Debrecen
[15] Michálek J.,Němcová M.,Popelínský L.,Řebíčková M. (1984) The system for multivariate statistical data processing . COMPSTAT 84,Proceedings in Computional Statist.
* 6th Symposium,Prague
[17] Řebíčková M.(1984) Využití log-lineárních modelů. Práce SVOČ, UJEP Brno
[18] Stoufer S.A.,Toby J.(1951) Role conflict and personality. Amer. J. Sociol. 56, 395-406
[19] Vajda I.(1982) Teória informacie a štatistického rozhodovania. ALFA Bratislava
[20] Whittaker J.(1982) GLIM syntax and simultaneous tests for graphical models. In Gilchrist R. Ed. GLIM 82 Springer Verlag, p. 98-108
[21] Whittaker J. (1984)Fitting all possible decomposable models to multidimensional contingency tables. COMPSTAT 84,Proceedings in Computional Statistics,6th Symposium
* Prague
[16] Rudas T. (1984) Testing goodness of fit of log-linear models based on small samples - a Monte Carlo study. Colloquium on goodness of fit,Debrecen