DISTANČNÍ PŘÍSTUP K ANALÝZE KATEGORIZOVANÝCH DAT A JEHO APLIKACE NA PROBLÉM SHODY

Jan ŘEHÁK (Ústav pro filosofii a sociologii ČSAV, Praha)

Blanka ŘEHÁKOVÁ (Ústav pro výzkum veřejného mínění při FSÚ, Praha)

Problém shody zahrnuje v analytické praxi širší třídu úloh, než jak je tradičně uváděn.

I. Úloha a motivace

Uveďme příklad čtyř kategorií $A = (A_1, A_2, A_3, A_4)$ s rovnoměrným rozdělením předpokládaných pravděpodobností výskytu

A_1	A_2	A_3	A_4	
.25	.25	.25	.25	teoretické rozdělení
.25	.25	.35	.15	první empirické rozdělení
.35	.25	.25	.15	druhé -"- -"-

V případě, že jde o prostou klasifikaci (nominální proměnnou), se první a druhé empirické rozdělení od teoretického liší ve zcela stejném stupni, pouze kvalita odchýlení (tj. kategorie, do niž se přesouvají jednotky z A_4) se mění. V případě, že jde o uspořádanou klasifikaci (v pořadí indexů), je už rozdíl velmi značný. Jestliže mezi A_k uvažujeme jinou relaci vzhledem ke které shodu hodnotíme, může být situace neshody ještě více diferencována. (A_k mohou být např. místa na mapě - Bratislava, Praha, Most, Litvínov tvoří názorný extrémní případ; kategorie mohou reprezentovat také určité typy, které se od sebe liší v různém stupni). Obecně složitější relace mezi A_k se vyskytují často ve společenských, biologických a lékařských vědách, v ekologické, diagnostické, typologické, dialektologické a jiné problematice).

Z příkladu plynou tři hlavní problémy:

a) Jak odlišit různé situace (relace mezi A_k) při testování shody?

b) Jak měřit neshodu?

c) Jak specifikovat neshodu?

Poslední problém se řeší pomocí testování rezidui (případ prosté klasifikace a případy, kdy klasifikace vzniká jako kombinace dvou, přičemž marginální distribuce jedné nebo obou klasifikací je známa viz B. Řeháková [1979]). Měření neshody bylo pro nominální proměnnou navrženo H. Weilerem [1966].

II. Weilerův koeficient neshody a jeho diskuse

Označíme

A :	$A_1, A_2, \dots, A_k$	kategorie,
π :	$\pi_1, \pi_2, \dots, \pi_k$	hypotetické pravděpodobnosti,
f :	$f_1, f_2, \dots, f_k$	empirické relativní četnosti,
p :	$p_1, p_2, \dots, p_k$	skutečné pravděpodobnosti.

Test hypotézy H_s : $\pi = p$ za předpokladu multinomického rozdělení provádíme např. pomocí Pearsonova X^2

$$X^2 = n \frac{\sum_i (\pi_i - f_i)^2}{\pi_i} \sim X^2_{k-1} \quad \text{za } H_s. \tag{1}$$

H. Weiler [1966] definoval koeficient neshody $\bar{\varphi}$ jako normalizaci $\sqrt{X^2}$ mezi 0 a 1

(2) $$\bar{\varphi} = \sqrt{\frac{\pi_{min}}{1-\pi_{min}} \sum_{i=1}^{K} \frac{(\pi_i - f_i)^2}{\pi_i}}$$

a podal jeho diskusi. Koeficient však má dvě základní nevýhody:

a) Je vhodný pro prostou klasifikaci a není zřejmé jeho rozšíření na obecnější typy proměnné.

b) Platí

(3) $$\bar{\varphi}^2 = \frac{\pi_{min}}{1-\pi_{min}} \left[\frac{(\pi_{min}-f_m)^2}{\pi_{min}} + \sum_{i \neq m} \frac{(\pi_i - f_i)^2}{\pi_i} \right]$$

kde m je index příslušný k π_{min} (je-li jediné). Z (3) plyne, že $\lim_{\pi_{min} \to 0} \bar{\varphi}^2 = f_m^2$ a tudíž příspěvek druhého členu v (3) je pro malá π_{min} mizivý oproti příspěvku v kategorii f_m. (Obdobné výsledky platí i v případě, kdy je více kategorií s minimálním π_{min}). V tomto případě tedy $\bar{\varphi}$ neodráží rovnoměrně neshodu ve všech kategoriích (obdobný nedostatek má i X^2).

III. Distanční model pro analýzu zobecněných kategorizovaných proměnných (D-model).

Výše uvedené aspekty se snažíme překonat pomocí distanční analýzy distribucí (J.Řehák, B.Řeháková [1979]). Označíme $A = \{A_1,\dots,A_K\}$ množinu hodnot kategorizace A, $f = \{f_1,\dots,f_K\}$ distribuci z K-rozměrného simplexu $\Theta_K = \{f: \sum f_i = 1, \ f_i \geq 0\}$. Dále označíme $D = \| d_{ij} \|$, typu $K \times K$ matici skórů vytvářejících typ proměnné, d_{ij} = skóry určující relace (A_i, A_j) a splňující podmínky $d_{ij} = d_{ji}$ a $d_{ii} = 0$. Zobecněnou proměnnou definujeme jako $A = \{A_1,\dots,A_K; \ D\}$.

Příklady:

a) $d_{ij} = 1$ pro $i \neq j$, pak A se nazývá <u>nominální (prostá kategorizace)</u>,

b) $d_{ij} = |i-j|$, pak A se nazývá <u>ordinální (uspořádaná kategorizace)</u>.

c) $d_{ij} = (x_i - x_j)^2$, kde x_i jsou čísla přiřazená kategoriím $(x_i = x(A_i))$, pak proměnnou nazýváme <u>kardinální (číselná kategorizace)</u>.

d) Jestliže existuje K vektorů s M složkami $x_k = (x_{k1},\dots,x_{kM})$, $k=1,\dots,K$ tak, že $d_{ij} = \sum_{m=1}^{M} (x_{im} - x_{jm})^2$, pak hovoříme o <u>M-dimenzionální metrické proměnné</u>; pro $M=2$ se tato proměnná nazývá <u>areální</u>.

Základní charakteristiky (definice) a vlastnosti

a) Zobecněná míra <u>rozptylu</u>

(4) $$\text{Genvar}\ f = f' D f$$

b) Míra <u>polohy</u>

(5) $$c = A_k \Rightarrow d_k^* = \min_j d_j, \quad \text{kde} \quad d_j = E\, d_{ij} = \sum f_i\, d_{ij}$$

c) Věta (<u>existence semimetriky</u>)

Pro $f, g \in \Theta_K$ platí

(6) $$D(f,g) = \sqrt{(f-g)' D (g-f)}$$

je semimetrika na Θ_K, jestliže $D^* = \| d_{ij}^* \|$, $d_{ij}^* = d_{ik} + d_{jk} - d_{ij}$, typu $(K-1)\times(K-1)$ je pozitivně semidefinitní. Je-li D^* pozitivně definitní, pak $D(f,g)$ je metrika na Θ_K. $D(f,g)$ nazýváme D-semimetrikou, resp. D-metrikou na Θ_K.

d) Věta (<u>rozklad zobecněné variance</u>)

Buď $f = \sum p_r f_{(r)}$, kde $p_r \in \Theta_R$, $f, f_{(r)} \in \Theta_K$, pak pro D tvořící semimetriku platí

(7) $$\text{Genvar}\ f = \sum p_r\, \text{Genvar}\ f_{(r)} + \frac{1}{2} \sum_r \sum_s p_r p_s D^2(f_{(r)}, f_{(s)}) = \sum p_r\, \text{Genvar}\ f_{(r)} + \sum p_r D^2(f_{(r)}, f)$$

Příklady:

	nominální	ordinální	kardinální
Genvar f	$1-\sum f_i^2$	$2\sum F_i(1-F_i)$	2var X
c	modální kat.	mediánová kat.	kategorie průměru
D(f,g)	$\sqrt{\sum (f_i-g_i)^2}$	$\sqrt{2\sum (F_i-G_i)^2}$	$\sqrt{2}\,\lvert\bar{x}_f-\bar{x}_g\rvert$
	(metrika)	(metrika)	(semimetrika)

kde F_i, G_i jsou distribuční funkce, $\bar{x}_f, \bar{x}_g$ jsou průměry.

Důsledky:

a) Nominální analýza odpovídá např. metodě CATANOVA (analýza rozptylu pro kategorizovaná data) zavedené v práci Light, Margolin [1971], aplikujeme-li (7) na nominální proměnnou.

b) Analýza kardinálních proměnných přechází k lineárnímu modelu a metodě nejmenších čtverců.

c) Model poskytuje přístup k analýze ordinálních dat, která je analogická k přístupu Cramér-von Miesesovy statistiky pro spojitá data (viz J.Řehák [1976]).

IV. Aplikace D-modelu na testování dobré shody (asymptotická teorie)

U testů dobré shody tvoří D přirozenou ztrátovou funkci odvozenou z typu kategorizace a tudíž poskytuje váhy pro odchylky od předpokladu. Testy dobré shody můžeme založit na statistice typu (8):

$$d^2 = D^2(\pi, f) \tag{8}$$

pro hypotézu $\pi = p$. Její významnost lze určit podle algoritmu AS 106 (viz Sheil, O'Muircheartaigh [1977]).

Druhou možností je převedení na χ^2 distribuci pomocí následujícího postupu. Položme $f^* = (f_1,\dots,f_{k-1})$, obdobně π^*, g^*. Existuje Q takové, že $D^* = QQ'$, $Q'Q = I_M$, kde I_M je jednotková matice typu $M\times M$, M = hodnost D^*. Platí $(f^*-\pi^*) = g^* \sim N(0,\Sigma)$ za platnosti H_s. Pro všechna $\pi_i > 0$ je rozložení regulární $(\sigma_{ij} = n^{-1}[\delta_{ij}\pi_i - \pi_i\pi_j])$. Odtud $h = (h_1,\dots,h_M) = g^* Q \sim N(0, Q'\Sigma Q)$ a tudíž

$$X_D^2 = h(Q'\Sigma Q)^{-1} h' = (f^* - \pi^*)Q(Q'\Sigma Q)^{-1} Q'(f^* - \pi^*)' \tag{9}$$

je za platnosti H_s rozděleno jako χ_M^2.
Pro ordinální D je $h = (F_1,\dots,F_{k-1})$, $\Sigma_o = Q'\Sigma Q$ je určena prvky $\sigma_{o,ij} = \Pi_i(1-\Pi_j)$ pro $i \le j$, kde Π_i je distribuční funkce k π.
Pro kardinální D je problém převeden na obvyklý z-test $X_D^2 = z^2 = n(\bar{x}_f - \bar{x}_\pi)^2 / \mathrm{var}(X/\pi)$.
Nominální případ vede na X^2 z (1).

V. Aplikace na měření neshody

Pro $A = \{A_1, A_2, \dots, A_k;\ D\}$, $\pi, p, f \in Q_k$ zavedla B.Řeháková [1980] koeficient neshody

$$\rho = \frac{D(\pi, p)}{\max_p D(\pi, p)} \tag{10}$$

V praxi používáme výběrový analog (odhad metodou maximální věrohodnosti)

$$R = \frac{D(\pi, f)}{\max_p D(\pi, f)} \tag{11}$$

Maximum ve jmenovateli (11) pro běžné typy proměnných (nominální, ordinální, kardinální) je

$$\sqrt{1 - 2\,\pi_{min} + \sum_{k=1}^{K} \pi_k^2}\,,$$

$$\sqrt{2 \max \left(\sum_{k=1}^{K-1} \Pi_k^2,\ \sum_{k=1}^{K-1} (1-\Pi_k)^2 \right)}\,, \quad \Pi_k = \sum_{i=1}^{k} \pi_i$$

$$\sqrt{2}\,\max (\bar{x}_\pi - x_{min},\ x_{max} - \bar{x}_\pi),$$

kde x_{min}, x_{max} jsou minimální a maximální skóry přiřazené kategoriím, $\bar{x}_\pi = \sum_k \pi_k x_k$.

Vlastnosti R :

a) $0 \leq R \leq 1$; $R = 0 \Leftrightarrow \pi = f$ pro D - metriku

$\pi = f \Rightarrow R = 0$ pro D - semimetriku

b) $R \overset{as.}{\sim} N(\rho, \frac{v(R)}{n})$, $v(R)$ jsou uvedeny v citované práci.

Asymptotická normalita umožňuje

a) konstrukci intervalů spolehlivosti $R \pm z_\alpha \sqrt{\frac{v(R)}{n}}$,

b) test $\rho = \rho_0$,

c) testy $\rho_1 = \rho_2 = \ldots = \rho_N$ a párová porovnání $\rho_i = \rho_j$ pro nezávislé výběry.

Poznámka : Koeficienty neshody i jejich testy našly praktickou aplikaci při vyhodnocování reprezentativity výzkumů veřejného mínění a při porovnávání distribucí s nezávislým standardem (např. shoda reality a záměru u působení časopisu, televize, rozhlasu ap.).

VI. Další aplikace D - modelu

1. Koeficienty asociace - speciální případ je Wallisovo τ i korelační poměr η^2 aplikovaný na kontingenční tabulku, též koeficient β pro ordinální proměnné (viz Řehák [1976]).
2. Koeficienty parciální asociace.
3. Analýza kategorizovaných dat ve schématech ANOVA (včetně testů homogenity pro tabulku RxS).
4. Seskupování řádků tabulky (např. vážená centroidní metoda-viz Řehák,Řeháková [1982]).
5. Klasické multidimenzionální škálování řádků tabulky.

Výhody D-modelu spočívají

a) v překonání obtíží, které byly diskutovány v úvodních částech referátu, při řešení všech komparačních úloh spojených s obecnými typy kategorizací;

b) v poskytnutí jednotného pohledu a jednotné interpretace odvozených měr a statistických testů;

c) v možnosti práce s obecnými typy kategorizací, ale i v jednoduché metodice pro analýzu ordinálních dat;

d) v jednoduchém heuristickém základu a v možnosti jednoduchých grafických reprezentací výsledků (eukleidovská reprezentace simplexu Q_K).

Literatura

1. Light,R.J., Margolin,B.H. [1971] : An Analysis of Variance for Categorical Data. Journal of American Statistical Association 66, 534-544.

2. Rao,R.C. [1978] : Lineární metody statistické indukce a jejich aplikace. Academia,Praha.

3. Řehák,J. [1976] : Základní deskriptivní míry pro rozložení ordinálních dat. Sociologický časopis 4, 416-431.

4. Řehák,J., Řeháková B. [1979] : Základní charakteristiky proměnných s konečným počtem hodnot a distanční analýza jejich rozložení. Sociologický časopis 2, 214-233.

5. Řehák,J. Řeháková B. [1981] : Analýza kategorizovaných dat v sociologii (nepublikováno).

6. Řeháková,B. [1979] : Statistické ověřování reprezentativity : testy dobré shody. Sociologický časopis 6, 615-629.

7. Řeháková,B. [1980] : Statistické ověřování reprezentativity : koeficienty neshody. Sociologický časopis 6, 612-627.

8. Weiler,H. [1966] : A Coefficient Measuring the Goodness of Fit. Technometrics 2,327-334.

9. Sheil J., O'Muircheartaigh I. [1977] : Algorithm AS 106, The Distribution of Non-negative Quadratic Forms in Normal Variables. Applied Statistics, 92-98.