Grafové smíšené interakční modely a vyhledávání modelu

Marta Horáková, ÚSEB ČSAV,
Květná 8, 603 65 Brno, ČSSR

Teorií grafových smíšených interakčních modelů se zabývali zejména Lauritzen, Wermuth /1989/ jako vlastní podtřídou hierarchických smíšených interakčních modelů zkoumaných Edwardsem /199?/

Předpokládejme, že analyzovaná data jsou realizací náhodného výběru obecné ze smíšeného rozložení pravděpodobností. Diskrétní veličiny přítomné v datech označme velkými písmeny za začátku abecedy /A,B,C,D,.../, spojité veličiny velkými písmeny z konce abecedy /Z,X,Y,U,.../. V grafovém zobrazení přiřaďme každé diskrétní veličině symbol ● označený jménem veličiny, každé spojité veličině symbol O označený jménem spojité veličiny. Symboly ● a O jsou vrcholy obyčejného grafu se dvěma typy vrcholů. Některé vrcholy jsou spojeny neorientovanou hranou, pokud mezi odpovídajícími náhodnými veličinami je předpokládána přímá závislost, zatímco vynechání hrany znamená přítomnost určité podmíněné nezávislosti veličin. Každému interakčnímu grafu G jednoznačně odpovídá grafový smíšený interakční model, a to takový model, jehož hustota CG-rozložení /Conditional Gaussian/ je markovská vzhledem ke grafu G. Označíme-li hustotu CG-rozložení vzhledem k součinové míře obvyklé čítací míry a Lebesqueovy míry $f(i,y) = \exp\{g(i)+h(i)^T.y+\frac{1}{2}.y^T.k(i).y\}$, kde i je p-rozměrný vektor označující hodnoty diskrétních veličin, y je q-rozměrný vektor označující hodnoty spojitých veličin, g(i), h(i), k(i) jsou diskrétní, lineární a kvadratické parametry CG-rozložení, g(i) je reálné číslo, h(i) je q-rozměrný vektor reálných čísel, k(i) je qxq-rozměrná symetrická pozitivně definitní matice koncentrace, označíme-li Δ množinu diskrétních a Γ množinu spojitých veličin, pak f(i,y) odpovídá grafovému smíšenému interakčnímu modelu s grafem $G=(\Delta\cup\Gamma, E(\Delta\cup\Gamma))$ právě tehdy, když existují reálné funkce $\lambda_d(i)$, $\eta_d(i)_X$, $\psi_d(i)_{XY}$ pro každé $d\subseteq\Delta$, $X\in\Gamma$, $Y\in\Gamma$, nazývané diskrétní, lineární a kvadratické interakce, takové, že platí:

i/ jednoznačnost: Pro každé $b\subseteq d$, $b\neq d$ je $\sum_{\{j:j_b=i_b\}}\lambda_d(j) = 0$, $\sum_{\{j:j_b=i_b\}}\eta_d(j)_X = 0$, $\sum_{\{j:j_b=i_b\}}\psi_d(j)_{XY} = 0$

ii/ korespondence s kanonickými parametry:

$$g(i) = \sum_{d\subseteq\Delta}\lambda_d(i),\quad h(i)_X = \sum_{d\subseteq\Delta}\eta_d(i)_X,\quad k(i)_{XY} = \sum_{d\subseteq\Delta}\psi_d(i)_{XY},$$

iii/ G-gibbsovská vlastnost:

$\lambda_d(i) = 0$ pokud neexistuje klika $c\in C_\Delta$, taková, že $d\subseteq c\cap\Delta$, kde C_Δ je množina klik v podgrafu $G(\Delta,E(\Delta))$,

$\eta_d(i)_X = 0$ pokud neexistuje klika $c\in C_{\Delta X}$, taková, že $d\subseteq c\cap\Delta$, $X\in\Gamma$, kde $C_{\Delta X}$ je množina klik v podgrafu $G(\Delta\cup\{X\},E(\Delta\cup\{X\}))$,

$\psi_d(i)_{XY} = 0$ pokud neexistuje klika $c\in C_{\Delta XY}$, taková, že $d\subseteq c\cap\Delta$, $X\in\Gamma$, $Y\in\Gamma$, kde $C_{\Delta XY}$ je množina klik v podgrafu $G(\Delta\cup\{X,Y\},E(\Delta\cup\{X,Y\}))$.

Ekvivalence G-gibbsovské a G-markovské vlastnosti je pro grafové smíšené interakční modely dokázána v Lauritzen, Wermuth /1989/. Interpretace je čitelná přímo z interakčního grafu: jsou-li $a\subseteq\Delta\cup\Gamma$, $b\subseteq\Delta\cup\Gamma$, v grafu separovány množinou $c\subseteq\Delta\cup\Gamma$, pak $a \perp b \mid c$.

Vyhledáváním modelu rozumíme situaci, kdy je třeba vysvětlit neznámou strukturu závislosti dat jedním nebo více alternativními modely na základě informace skryté v datech, aniž jsou nějaké hypotézy stanovené Havránek/1982/, Edwards, Havránek/1985/.

Zkoumání struktury závislosti pouze diskrétních náhodných veličin odpovídá studiu logaritmicko-lineárních interakčních modelů /např. Bishop a kol. 1975/. Pokud data obsahují pouze spojité veličiny, jedná se o problematiku výběru struktury kovarianční matice /Dempster 1972, Wermuth 1980, 1986/. Při výběru modelů struktury závislosti v čistě diskrétním a v čistě spojitém případě se vedle jiných přístupů /Wermuth a kol. 1976, Edwards 1984, Whittaker 1984/ hodně používaly algoritmy využívající částečné uspořádání modelů dané inkluzí modelů a některá pravidla zabezpečující redukci počtu přímo testovaných modelů. Ve spojitosti s optimálním výběrem regresorů použil např. Beale a kol. /1967/ pravidlo, které označíme PA a které znamená automatické akceptování bez přímého testování všech modelů „větších" /vzhledem k částečnému uspořádání daného inkluzí/ než model akceptovaný /nezamítnutý/ daným testem přímo. Ve spojitosti s výběrem modelů logaritmicko-lineárních grafových se naopak používalo pravidlo PR /Havránek 1982/, které předpokládá automatické zamítnutí bez přímého testování všech modelů „menších" než model zamítnutý přímým testem. Kombinace obou pravidel je poprvé použita pro analýzu vícerozměrných kontingenčních tabulek v Edwards, Havránek /1985/. Na obecnou třídu modelů s jistými algebraickými vlastnostmi byl postup s Konstrukcí A-duálu a R-duálu množin modelů zobecněn v Edwards, Havránek /1987/.

Zkoumejme tedy algebraické vlastnosti třídy grafových smíšených interakčních modelů. Nechť $\mathcal{G}$ je třída grafových smíšených interakčních modelů s p diskrétními a q spojitými veličinami. Nechť $M \in \mathcal{G}$, $L \in \mathcal{G}$, nechť $G_M = (\Delta \cup \Gamma, E_M)$ je graf modelu M, $G_L = (\Delta \cup \Gamma, E_L)$ je graf modelu L. Řekneme, že $M \leq L$ právě tehdy, když $E_M \subseteq E_L$. Třída $(\mathcal{G}, \leq)$ je konečný svaz. Definujme dále operace průsek $(\wedge)$, spojení $(\vee)$ a komplement $(')$ přirozeným způsobem: $M \wedge L = K$ právě tehdy, když $G_K = (\Delta \cup \Gamma, E_K)$ a $E_K = E_M \cap E_L$,

$M \vee L = K$ právě tehdy, když $G_K = (\Delta \cup \Gamma, E_K)$ a $E_K = E_M \cup E_L$,

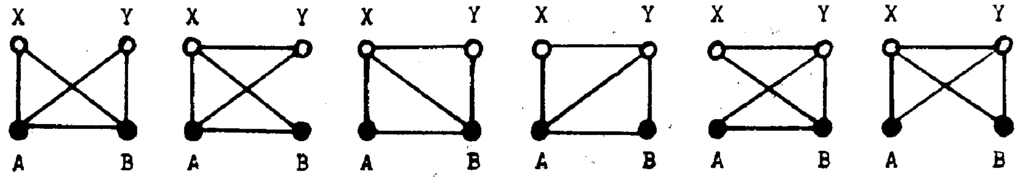
$M' = K$ právě tehdy, když $G_K = (\Delta \cup \Gamma, E_K)$ a $E_K = \bar{E}_M$,

kde $\cap$, $\cup$, $\bar{\ }$ jsou běžné množinové operace průnik, sjednocení a doplněk. $(\mathcal{G}, \wedge, \vee, ')$ je konečná Booleova algebra. Vlastnosti konečných Booleových algeber se projeví v algoritmech pro konstrukci A-duálu a R-duálu množiny modelů /Edwards, Havránek 1987/. Z hlediska výpočetního jsou podrobně popsány v Horáková /1989/. Jejich počítačová realizace na IBM PC AT kompatibilních počítačích je začleněna v systému MIMAS jako rozšíření Edwardsova systému MIM /Edwards 1987,1989/ o automatické vyhledávání modelů. Rozšíření je provedeno zcela v duchu konstrukce systému MIM. Syntax a význam příkazových slov CRITLEVEL, REPORT, AUTOSEARCH, INITSEARCH, STARTSEARCH spojených s automatickým výběrem je vysvětlena v systému nápovědy systému MIM-MIMAS. Příkaz REPORT ovlivňuje úroveň výstupní informace o průběhu automatického vyhledávání, bez volby REPORT se vypíše jen informace o počtu přímo testovaných modelů a informace o adekvátních grafových smíšených interakčních modelech, které jsou vlastně $1-\alpha$ množinou spolehlivosti /Havránek, Soudský 1989/ pro model popisující strukturu závislosti analyzovaných dat. Příkaz CRITLEVEL umožní měnit hladinu významnosti používanou v automatickém vyhledávání. Příkaz AUTOSEARCH spouští proces vyhledávání, INITSEARCH umožňuje navíc omezit vyhledávání pouze na submodely určitého modelu nebo na modely, které obsahují specifikovaný model jako submodel. STARTSEARCH zase umožňuje ukončit vyhledávání jakmile počet přímo testovaných modelů překročí udanou hodnotu.

Ilustraci postupu automatického vahledávání ukazuje příklad použití na datech z Beklová a kol. /1988/ se dvěma diskrétními náhodnými veličinami /A = druh s hodnotami: 1 = Phasianus colchicus "hybrid" světlý, 2= Phasianus colchicus "hybrid" tmavý, 3= Phasianus colchicus colchicus f. tenebrosus, B = pohlaví shodnotami: 1=samec, 2= samice/ a dvěma spojitými veličinami /X=délka pravého křídla/v mm/, Y=délka pravého tarsometatarsu /v mm//. Zadáním příkazů FACTOR A3B2; CONTINUOUS XY definujícím diskrétní a spojité proměnné, načtením vstupních dat např. příkazem CELLREAD ABXY , volbou REPORT a a CRITLEVEL 0.05 a zadáním AUTOSEARCH je v systému MIMAS zahájen proces automatického vyhledávání pro daná čtyřrozměrná data. Přímo je testováno pouze 7 modelů, průběh testování je patrný z obrázku.

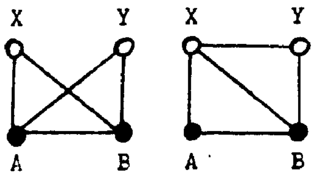
Obrázek:

Počáteční množina testovaných modelů:

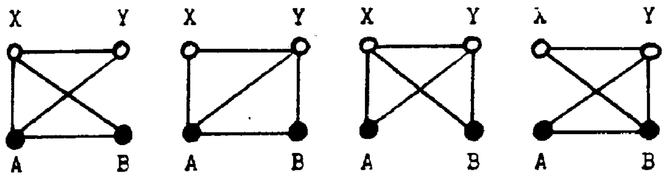


LR =	10.5940	129.0368	14.2603	288.6336	39.4798	28.3488
DF =	6	9	4	9	12	12
P =	0.1010	0.0	0.2837	0.0	0.0002	0.0052
	akceptován	zamítnut	akceptován	zamítnut	zamítnut	zamítnut

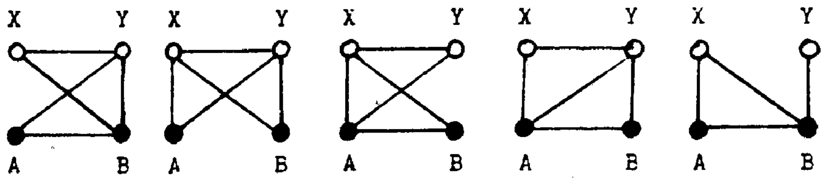
množina a akceptovaných modelů:



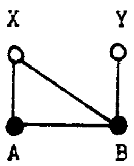
množina zamítnutých modelů $\bar{a}$:



$D_R(a)$:



$D_A(a)$:



LR= 21.9400 interpretace: A,X $\perp$ Y | B .

DF= 14

P = 0.0795

$D_R(a) - \bar{a} = D_A(a) - \bar{a} = D_A(a)$: výsledný model.

Literatura:

Beale, E., Kendall,M., Mann,D., 1967: The discarding of variables in multivariate analysis. Biometrika 54, str. 357-365.

Beklová,M., Hanák,V., Pikula, J., 1988: Morphometry of Phasianus colchicus during the hunting season. Acta Sc. Nat. Brno 22/11/, str. 1-64.

Bishop, Y.M.M., Fienberg,S.E., Holland, P.W., 1975: Discrete multivariate analysis: Theory and practice. MIT Press.

Dempster, A.P.D., 1972: Covariance selction. Biometrics 28, str. 157-175.

Edwards,D., 1984: A computer intensive approach to the analysis of sparse multidimensional contingency tables. COMPSTAT'84. Physica Verlag, Wien, 335-359.

Edwards,D., 1990: Hierarchical interaction models. JRSS Serie B, 52/1/, str 3-20.

Edwards, D., 1987,1989: A guide to MIM. Výzkumné zprávy, Universita Kodaň.

Edwards,D., Havránek, T., 1985: Model search in contingency tables. Biometrika 72, str 339-351.

Edwards, D., Havránek, T., 1987: A fast model selection procedure for large families of models. Jour. of Amer. Stat. Ass. 82, str. 205-213.

Havránek, T., 1982: O analýze mnohorozměrných kontingenčních tabulek. ROBUST´82, str.11-16.

Havránek, T., Soudský, C., 1989: Model choice in the context of simultaneous inference. In: Dodge, Y., /ed./: Statistical data analysis and inference. North Holland. str. 165-176.

Horáková, M., 1988: Smíšené interakční modely. ROBUST´88, str.41-44.

Horáková, M., 1989: Dependence structure selection using graphical mixed interaction models. Určeno pro CSD .

Lauritzen, S.L., Wermuth, N., 1989: Graphical models for associations between variables, some of which are qualitative and some quantitative. The Annals of Stat. 1, str. 31-57.

Wermuth, N., Wehner, T., Gönner, H., 1976: Finding condensed descriptions for multidimensional data. Computer programmes in biomedicine 6, str 23-38.

Wermuth, N., 1980: Linear recursive equations, covariance selection and path analysis. Jour. of Amer. Stat. Assoc. 75, str. 963-972.

Wermuth, N., 1986: Aspects of different parametrizations of a join normal distribution. Výzkumná zpráva č.2, Universita Mnichv.

Whittaker, J., 1984: Fitting all possible decomposable models to multidimensional contingency tables. COMPSTAT´84. Physica Verlag, Wien.