Tomáš Havránek

Automatické formování hypotéz metodou GUHA - teorie a aplikace

Pokroky matematiky, fyziky a astronomie, Vol. 26 (1981), No. 3, 136--150,151

Persistent URL: http://dml.cz/dmlcz/138862

obrázek 35 z Bradford Washburn, obrázek 36 z U.S. Naval Research Laboratory, obrázek 39 z California Institute of Technology, obrázek 7 z Stanford Linear Accelerator Center a obrázek 33 z U.S. Aerospace Corporation.

Rád bych poděkoval Sandře Sicignano za její pomoc při získávání obrázků, Frances de Grenier a Kláře Buckleyové za přepisování rukopisu a nejvíce Marii Anně Thomsonové-Schulzové za její pomoc při přípravě tohoto článku. Nakonec bych rád vyjádřil své poděkování nadaci Alexandr von Humboldta za udělení Humboldtova stipendia na dobu, kdy jsem připravoval tento článek.

Automatické formování hypotéz metodou GUHA — teorie a aplikace

Tomáš Havránek, Praha

Metoda GUHA (General Unary Hypotheses Automaton) se objevila ve své základní verzi před patnácti lety. Byla dílem tří českých matematiků — Petra Hájka, Ivana Havla a Metoděje Chytila. Od té doby prošla dosti bouřlivým rozvojem, práce na ní se zúčastnilo několik dalších spolupracovníků a měla poměrně značné úspěchy (hlavně po teoretické stránce). Vzbudila zájem i v zahraničí, je jí například věnováno celé první číslo ročníku 1978 časopisu International Journal of Man-Machine Studies. Vzbudila i dosti diskusí a pochybností. Pokusíme se nyní ohlédnout za jejím vývojem a shrnout její dnešní stav.

Tento článek vznikl na přání redakce Pokroků sloučením dvou textů: jednak textu přednášky přednesené na konferenci českých matematiků ve Zvíkovském podhradí v roce 1978 a publikovaném ve sborníku z této konference, jednak z původního textu připravovaného pro Pokroky, jehož napsání bylo v přednášce slibováno. Tím je vysvětleno to, že zde čtenář může nalézt formulace již otištěné jinde.

I. Úvod

1.1 Metoda GUHA je v podstatě pokusem aplikovat vyjadřovací a deduktivní prostředky matematické logiky na *analýzu empirických dat*. Základní *aplikační situace*, o kterou zde jde, je tato:

Máme před sebou data, která jsou výsledkem jisté *observační studie*; to znamená, že byla získána pozorováním řady vlastností a veličin na zpravidla velkém počtu objektů (například pozorováním různých anamnestických a diagnostických údajů na velkém množství pacientů). Od *experimentální studie* se naše situace liší v tom, že v experi-

mentální studii zpravidla máme již dánu jistou poměrně specifickou hypotézu – partikulární cíl výzkumu (např. že preparát xy kladně ovlivňuje průběh choroby A) a pro *ověření* této hypotézy *uspořádáme experiment* (tj. v našem příkladě vyberem pokusnou a kontrolní skupinu pacientů a zkoumáme, zda vskutku bude průběh choroby A u pokusných pacientů ošetřených preparátem xy příznivější než u pacientů kontrolních – pomiňme ovšem nyní etickou stránku klinických experimentů).

V observační studii známe na počátku zpracování dat pouze globální a zpravidla více či méně vágní cíl výzkumu: objevit například příčiny komplikací průběhu choroby A. Pro tento účel pozoroval výzkumník celou řadu pacientů postižených touto chorobou a zaznamenával pokud možno „všechny" jejich vlastnosti a „všechny" informace o průběhu choroby. Co je to všechno, je určeno obvykle intuicí a zkušeností výzkumníka, proto se v observačních datech zcela zákonitě objevuje i velké množství informačního balastu.

Jiné rozdělení úloh analýzy dat je na úlohy konfirmační a explorační. Konfirmační úlohy jsou většinou experimentální a mají rysy, o kterých jsme v souvislosti s experimentálními studiemi mluvili výše. Jejich řešení je věcí klasické matematické statistiky. Explorační úlohy bývají velmi často observačního charakteru. Naše aplikační situace je tedy přesněji řečeno explorační i observační. Metoda GUHA se může týkat i zpracování experimentálních dat v explorační úloze.

Budeme studovat struktury tohoto typu: je dána konečná množina M objektů a konečný počet vlastností $V_1, \ldots, V_n$ těchto objektů (v praxi bývá počet prvků množiny M řádově $10^2 - 10^3$ a počet vlastností řádově $10 - 10^2$). Každou vlastnost V_i „ohodnocujeme" na množině M pomocí funkce $\|V_i\| : M \to \mathscr{V}_i$, kde $\mathscr{V}_i$ je některá množina racionálních čísel. Speciálním případem je, když $\mathscr{V}_i = \{0, 1\}$ pro $i = 1, \ldots, n$ (jde o dvouhodnotová data získaná zjišťováním nepřítomnosti nebo přítomnosti nějaké vlastnosti) nebo $\mathscr{V}_i = \{0, \ldots, k_i\}$, kde k_i je předem dané malé přirozené číslo.

Z formálního hlediska jde o strukturu („soubor dat") $\boldsymbol{M} = \langle M, \|V_1\|, \ldots, \|V_n\| \rangle$. (Pro zjednodušení zde píšeme $\|V_i\|$ místo korektního $\|V_i\|_M$.)

1.2 Data takového typu je třeba analyzovat. Je jasné, že ke globálnímu cíli výzkumu existuje velká řada relevantních otázek typu „ověření" jednoduché partikulární hypotézy; jako např. zda vlastnost V_i souvisí s vlastností V_j, $i \neq j$, tj. zda kladně ovlivňuje přítomnost vlastnosti V_j. Otázek tohoto typu zpravidla výzkumník vymyslí celou řadu. Takto získaný seznam otázek je však nutně neúplný, neboť není v lidských silách „všechny" relevantní a formálně přípustné otázky vyslovit; uvědomme si jen, že nás může zajímat $\binom{n}{2}$ prostých binárních vztahů mezi vlastnostmi popsanými výše.

1.3 Je zde tedy tento problém: Jak vytvořit v jistém smyslu úplný seznam otázek, které mají být kladeny? Jak úsporně popsat data? Jak z nich „vytáhnout" relevantní informace? atd.

Je zde zřejmá potřeba automatizace: zadat snadno a rychle, které otázky klást, automatizovat jejich vytváření a zodpovídání.

Skrývá se tu však veliké nebezpečí: čas a prostor, který bude potřebovat stroj k vytvoření a zodpovědění otázek, a čas a prostor, který bude potřebovat člověk k prostudování

odpovědí. Ten, kdo měl v ruce strojové výpisy výsledků některých programů pro analýzu dat, jistě pochopí, že i prostor zde hraje pro člověka roli.

Abychom dokázali zodpovědět otázky položené výše a vyhnout se nebezpečí, které na nás číhá, je nutné vytvořit precizní popis celé situace; vytvořit formální aparát, ve kterém by bylo možné některé problémy přesně popsat a vyřešit. Konečně je patrně zřejmé, že formální popis je conditio sine qua non pro automatizaci (computerizaci) celého postupu analýzy dat.

II. Metoda GUHA

2.1 Jeden z možných přístupů v případech výše uvedených vytváří metoda GUHA, respektive její matematické prostředky v souvislosti s ní vytvořené. V metodě GUHA jde vždy primárně o popis dat pomocí sentencí (tvrzení) určitého předem popsaného tvaru. Tyto sentence jsou automaticky generovány, vyhodnocovány a je tištěn seznam pravdivých sentencí v jisté úsporné formě. O jaké sentence jde, si vysvětlíme postupně na příkladě jednoduchých procedur metody GUHA.

Historickým začátkem metody GUHA jsou práce [5] a [6]. Zkoumaná data byla v těchto prvních pracích jednoduchá – dvouhodnotová, tj. $\mathscr{V}_i = \{0, 1\}$, $i = 1, \ldots, n$. Šlo tedy o rozbor matic nul a jedniček, z věcného hlediska například o výsledky dotazníkových šetření s odpověďmi ne – 0, ano – 1. Zde se nabízí „přirozený" prostředek popisu – predikátový počet. Zkoumáme n vlastností – predikátů $V_1, \ldots, V_n$. V souhlase s dříve zavedenou symbolikou, jestliže $o \in M$ je nějaký objekt, pak klademe $\|V_i\|\,[o] = 1$ nebo 0 podle toho, zda objekt o má nebo nemá vlastnost V_i. Pomocí obvyklých logických spojek & (konjunkce, a), $\vee$ (disjunkce, nebo) a $-$ (negace, ne) můžeme z vlastností $V_1, \ldots, V_n$ vytvářet nové, *derivované* vlastnosti. Například $V_1 \,\&\, V_2$ je vlastnost, která spočívá v současné přítomnosti vlastnosti V_1 i vlastnosti V_2 : $\|V_1 \,\&\, V_2\|\,[o] = 1$, jestliže $\|V_1\|\,[o] = 1$ i $\|V_2\|\,[o] = 1$. Objekt o má vlastnost $V_1 \,\&\, V_2$, má-li obě vlastnosti V_1 i V_2. Podobně $V_1 \,\&\, -V_2$ znamená přítomnost vlastnosti V_1 a nepřítomnost vlastnosti V_2. Derivované vlastnosti tvaru $V_1 \& - V_2 \& V_7$ atd., kde se žádná vlastnost nevyskytuje více než jednou, se nazývají *elementární konjunkce*, zkratka EK. ($V_1 \& - V_1$ není elementární konjunkce.) Podobně formule tvaru $V_1 \vee V_2 \vee V_6 \vee - V_7$ apod. se nazývají *elementární diskunkce*, zkratka ED.

Derivované vlastnosti označujeme písmeny $\varphi, \psi, \lambda, \ldots$ Jsou-li φ a ψ dvě derivované vlastnosti, pak z původní struktury $\boldsymbol{M} = \langle M, \|V_1\|, \ldots, \|V_n\| \rangle$ obdržíme *derivovanou strukturu* $\boldsymbol{M}_{\varphi,\psi} = \langle M, \|\varphi\|, \|\psi\| \rangle$.

2.2 Čtenář jistě zná další výrazový prostředek klasického predikátového počtu – *kvantifikátory*. Klasické kvantifikátory jsou dva $\forall$ a $\exists$ („pro všechny" a „existuje"). Můžeme si položit otázku „$\forall V_1 \vee -V_2$"?. Je pravda, že každý objekt má vlastnost V_1 nebo nemá vlastnost V_2?

V naší konečné situaci umíme na tuto otázku odpovědět: $\forall V_1 \vee -V_2$ je pravda v $\boldsymbol{M}$ (tj. $\|\forall V_1 \vee -V_2\|_{\boldsymbol{M}} = 1$), jestliže pro každé $o \in M$ je $\|V_1 \vee -V_2\|\,[o] = 1$. Odpověď dostáváme prozkoumáním hodnoty na každém objektu $o \in M$ a to je pro konečnou mno-

žinu objektů proveditelné. Pro konečný soubor dat jsou tedy klasické kvantifikátory *efektivně* vyhodnotitelné. To je tedy podstatný rozdíl od nekonečných struktur zkoumaných v matematické logice obecně.

2.3 Jistě už na tomto bodu výkladu umíme snadno popsat jistou třídu vícemeně zajímavých otázek: $RQ = \{\forall\varphi \mid \varphi \text{ je ED}\}$ – ptáme se, zda některá elementární disjunkce neplatí pro všechny objekty z M (tj. nepopisuje v jistém smyslu úplně data). Pokusíme se formulovat jiné otázky, které mají sémanticky jasnější smysl: např. „$V_1 \Rightarrow -V_2$"? Je tomu tak, že vždy, když objekt o má vlastnost V_1, pak nemá vlastnost V_2? $V_1 \Rightarrow -V_2$ bude pravdivé v datech $\boldsymbol{M}$ (tj. $\|V_1 \Rightarrow -V_2\|_{\boldsymbol{M}} = 1$), když pro každé $o \in M$ je $\|V_1\|[o] \leqq \leqq \|-V_2\|[o]$. Vidíme, že pro ohodnocení $V_1 \Rightarrow -V_2$ použijeme pouze strukturu $\boldsymbol{M}_{V_1,-V_2} = \langle M, \|V_1\|, \|-V_2\|\rangle$. Podobně pro obecné derivované vlastnosti φ a ψ ohodnocujeme formuli $\varphi \Rightarrow \psi$ ve struktuře $M_{\varphi,\psi} = \langle M, \|\varphi\|, \|\psi\|\rangle$. Co je společné formulím $V_1 \Rightarrow -V_2$ a $\forall V_1 \vee -V_2$: jejich hodnota nezávisí na jednotlivých objektech, ale je přiřazována celé struktuře $\boldsymbol{M}$. Takové formule nazýváme *sentencemi*. Co má společného $\Rightarrow$ a $\forall$? Jsou to oba kvantifikátory: $\Rightarrow$ je *zobecněný kvantifikátor*. Pomocí $\Rightarrow$ i $\forall$ vytváříme z derivovaných vlastností sentence. Pouze sentence odpovídají otázkám, ve kterých se ptáme na „kvalitu" celé struktury dat.

Kvantifikátor $\Rightarrow$ je ovšem jednoduchý kvantifikátor v tom smyslu, že ho lze *definovat* pomocí obvyklých spojek $\vee$ a $-$ a pomocí klasického obecného kvantifikátoru $\forall$. Sentence $\varphi \Rightarrow \psi$ je totiž pravdivá, právě když je pravdivá sentence $\forall(-\varphi \vee \psi)$. V tomto smyslu $\Rightarrow$ není podstatným zobecněním predikátového počtu.

Jistě lze pomocí $\Rightarrow$ popsat celou řadu rozumných otázek, např.: $RQ = \{\varphi \Rightarrow \psi \mid \varphi \in \in A, \psi \in B, \varphi \text{ a } \psi \text{ neobsahují společnou vlastnost}\}$, kde A, B jsou nějaké snadno popsatelné množiny elementárních konjunkcí, tj. lze je zadat jednoduchými syntaktickými pravidly, např. v A jsou všechny elementární konjunkce vytvořené z $V_1, \ldots, V_{17}$, a to takové, že neobsahují více než tři vlastnosti současně. B může být jednoduché, např. $B = \{V_{20}\}$. Podle tohoto návodu by jistě již každý dokázal *generovat* množinu otázek RQ. Věcný smysl našich otázek je tento: ptáme se, za jakých okolností, vzniklých kombinováním $V_1, \ldots, V_{17}$, vždy nastalo V_{20}.

Z praktického hlediska jsou to příliš „silné" otázky – může nás zajímat, kdy nastalo pouze „skoro" vždy V_{20}.

2.4 Tím se dostáváme k řadě otázek vytvořených kvantifikátory *statisticky inspirovanými*. Jsou-li φ a ψ dvě derivované vlastnosti, které jsou disjunktní, tj. neobsahují žádnou společnou vlastnost, a $\boldsymbol{M}$ data, označíme

$$a = \operatorname{card}\{o \in M \mid \|\varphi \,\&\, \psi\|[o] = 1\}, \qquad b = \operatorname{card}\{o \in M \mid \|\varphi \,\&\, -\psi\|[o] = 1\},$$
$$c = \operatorname{card}\{o \in M \mid \|-\varphi \,\&\, \psi\|[o] = 1\}, \quad d = \operatorname{card}\{o \in M \mid \|-\varphi \,\&\, -\psi\|[o] = 1\}.$$

Číslo a je tedy počet objektů, které mají současně obě derivované vlastnosti φ, ψ atd. Označme si $r = a + b$, $s = c + d$, $k = a + c$, $l = c + d$. Zřejmě $m = a + b + c + d$. Můžeme pak tvořit otázky typu: „$\varphi \Rightarrow_p \psi$?" To znamená, že se ptáme, zda za okolností φ nastala situace ψ s relativní četností větší nebo rovnou p, kde p je blízké jedničce, např. $p \geqq 0{,}95$. V naší symbolice: $\|\varphi \Rightarrow_p \psi\|_M = 1$, jestliže relativní četnost $a/(a + b)$ je $\geqq p$.

2.5 Jiný typ otázek chápe úlohu φ a ψ symetricky, ptáme se, zda φ a ψ kladně souvisí „$\varphi \sim {}^{\circ}\psi$“?. Zde $\|\varphi \sim {}^{\circ}\psi\|_M = 1$, jestliže $ad > bc$ – počet shodných nastání nebo nenastání obou vlastností převyšuje počet neshod. To je ovšem velmi primitivní otázka, odpověď na ni neobsahuje mnoho užitečné informace. Jiná otázka je, zda spolu φ a ψ „*významně*“ souvisí: „$\varphi \sim {}_{\alpha}\psi$“?, kde $\alpha \in (0; 0{,}5)$. Sémantika: $\|\varphi \sim {}_{\alpha}\psi\|_M = 1$, jestliže $ad > bc$ a zároveň

$$\sum_{i=a}^{\min(r,k)} \sigma(i, r, k, m) \leqq \alpha ,$$

kde $\sigma(i, r, k, m) = (r!\, s!\, k!\, l!)/(i!(r-i)!\,(k-i)!\,(m+i-r-k)!\, m!)$. Ve vymýšlení podobných otázek a je vytvářejících kvantifikátorů se meze nekladou [9].

2.6 Co je důležité pro tyto kvantifikátory:

a) Důležitá je jejich statistická interpretace. Při jejich konstrukci vycházíme z jistých předpokladů o vzniku dat: předpokládáme, že data vznikají například nezávislým mnohorozměrným náhodným výběrem nebo experimentem; ohodnocení vlastností na jednotlivých objektech je tedy výsledkem realizace téže n-rozměrné náhodné veličiny a navíc realizace na objektech jsou stochasticky nezávislé – pro každou k-tici objektů jde o vektor k n-rozměrných náhodných veličin s týmž rozložením pravděpodobností. Kromě toho se zpravidla předpokládá, že pro každou n-tici nul a jedniček je pravděpodobnost jejího výskytu u každého objektu nenulová. Pak můžeme mluvit o pravděpodobnosti nastání – přítomnosti – vlastností φ, ψ, $\varphi \,\&\, \psi$ atd. Z kladné odpovědi na otázku „$\varphi \Rightarrow {}_{p}\psi$“? pak můžeme například usuzovat na to, že $P(\varphi \,\&\, \psi)/P(\varphi) \geqq p$, tj. podmíněná pravděpodobnost nastání ψ za předpokladu nastání φ je větší nebo rovna p. (musíme si ovšem být vědomi stupně oprávněnosti takového usuzování). Podobně z $\|\varphi \sim {}_{\alpha}\psi\|_M = 1$ usuzujeme, že obecně (v celé populaci, jejíž vzorek zkoumáme) je $P(\varphi \,\&\, \psi)\, P(-\varphi \,\&\, -\psi) > P(\varphi \,\&\, -\psi)\, P(-\varphi \,\&\, \psi)$; φ a ψ spolu tedy obecně kladně souvisí (zde je oprávněnost usuzování v jistém smyslu vyšší než v předchozím případě, a to vzhledem ke konstrukci kvantifikátoru na základě testové statistiky).

Shrňme: konstrukce je svázána s představou o původu dat, o mechanismu vzniku dat.

b) Kvantifikátory výše uvažované jsou *klasicky nedefinovatelné* – nedají se vyjádřit pomocí $\forall, \exists, \&, \vee, -$. Znamená to, že vytvářejí skutečné zobecnění výrazových prostředků predikátového počtu.

c) Kvantifikátory lze shrnout do určitých tříd. Například $\Rightarrow, \Rightarrow_p$ patří do třídy implikačních kvantifikátorů, pro které je charakteristická tato vlastnost: Pro každé $\boldsymbol{M}$, je-li $a_{\varphi,\psi} \geqq a_{\lambda,\kappa}$ a $b_{\varphi,\psi} \leqq b_{\lambda,\kappa}$ a $\|\varphi \to \psi\|_M = 1$, pak i $\|\lambda \to \kappa\|_M = 1$. Zde φ, ψ a λ, κ jsou páry disjunktních derivovaných vlastností a $a_{\varphi,\psi}, \ldots$ atd. jsou frekvence vztahující se k těmto párům (viz. 2.4) a $\to$ značí libovolný implikační kvantifikátor, tj. např. $\Rightarrow$ nebo $\Rightarrow_p$. Konkrétní algoritmické procedury pro generování a zodpovídání otázek jsou konstruovány pro celou třídu kvantifikátorů najednou – mění se pouze konkrétní způsob evaluace kvantifikátoru.

2.7 Podstatné je využití dedukčních pravidel korektních vždy pro celou třídu kvantifikátorů. Nechť nyní je $\to$ libovolný implikační kvantifikátor. Pak je korektní následující

dedukční pravidlo: pro libovolné páry φ, ψ a λ, κ, kde φ, ψ jsou EK, λ, κ jsou ED a $\varphi, \psi, \lambda, \kappa$ jsou disjunktní; z platnosti $\varphi \,\&\, \psi \to \lambda$ plyne platnost $\varphi \to \lambda \vee \kappa \vee -\psi$. V logické symbolice zapisujeme toto dedukční pravidlo ve tvaru:

$$I_1 \qquad \frac{\varphi \,\&\, \psi \to \lambda}{\varphi \to \lambda \vee \kappa \vee -\psi} .$$

Označme symbolem $\Leftrightarrow$ obvyklou logickou ekvivalenci (levá strana je pravdivá, právě když je pravdivá pravá strana) a $\varphi \leqq \varphi'$ nechť označuje, že konjunkce φ je obsažena v konjunkci φ'. Potom za hořejších předpokladů je pro každé φ' takové, že $\varphi \leqq \varphi' \leqq \psi$, následující dedukční pravidlo korektní:

$$I_2 \qquad \frac{\varphi \to \lambda,\ \varphi \Leftrightarrow \psi}{\varphi' \to \lambda} .$$

I_1 je pravidlo *přímé*, obsahuje pouze sentence, které jsou našimi relevantními otázkami, I_2 je pravidlo *nepřímé*, obsahuje pomocné sentence, které nejsou předmětem našeho zájmu. Užitečnost pravidel je zřejmá: z platnosti jedné nebo dvou sentencí vidíme v jediném kroku platnost celé řady dalších sentencí (otázek). Tuto dedukci na první pohled použijeme k redukci procedurou tištěných kladných odpovědí.

2.8 Nyní si popíšeme úkol tzv. „implikační verze" metody GUHA. Uvažujme množinu otázek

$$RQ = \{\varphi \to \psi \mid \varphi \in A \subseteq \text{EK}, \psi \in B \subseteq \text{ED}, \varphi, \psi \text{ disjunktní}\} .$$

Množiny A, B jsou zde snadno syntakticky popsatelné (viz 2.3). Úkol pro počítačovou proceduru zní: je-li dán soubor dat $\boldsymbol{M}$ a množina otázek RQ, je nutné nalézt množinu X sentencí, kterou nazýváme *řešením*, a to takovou, aby

(1) každá sentence z X byla pravdivá v $\boldsymbol{M}$,

(2) každá pravdivá sentence z RQ (v $\boldsymbol{M}$) byla buď obsažena v X, nebo bezprostředně vyplývala ze sentencí obsažených v X použitím dedukčních pravidel I_1 a I_2.

To je obecná formulace úkolu metody GUHA, odmyslíme-li si speciální zadání RQ a I výše.

Cílem je ovšem nalézt X co *nejmenší* – nalézt úsporný popis dat. V našem speciálním implikačním případě je zkonstruována procedura, která v případě použití pouze $I = I_1$ dává řešení $X \subseteq RQ$, které je kardinálně minimální; jde tedy o řešení s minimálním počtem prvků. V případě $I = I_1 \cup I_2$ dostáváme řešení inkluzivně minimální, tj. řešení, jehož žádná vlastní část není řešením.

Podobně je možno popsat i „*asociační* verzi" metody GUHA, kdy

$$RQ = \{\varphi \sim \psi \mid \varphi \in A \subseteq \text{EK}, \psi \in B \subseteq \text{EK}, \varphi, \psi \text{ disjunktní}\}$$

a kde $\sim$ je kvantifikátor „souvisení" – asociační.

2.9 Podstatnou věcí je, aby procedura *nepracovala exhaustivním hledáním*, aby se neptala na platnost každé otázky z RQ, resp. aby negenerovala všechny otázky z RQ. Procedura generuje otázky v lineárním uspořádání $q_1 < q_2 < \ldots < q_n < \ldots$ (obvykle jde o uspořádání podle počtu obsažených vlastností a lexikografické). Kritickým bodem realizace procedury je možnost nalezení pozitivních a negativních skoků: možnost z platnosti (či neplatnosti) např. q_i usoudit na platnost (či neplatnost) otázek z celého úseku $q_{g(i)} < \ldots < q_{h(i)}$ a otázky z tohoto úseku pak již negenerovat (přitom stanovení $g(i)$ a $h(i)$ musí být jednoduché). Tato otázka je řešena pozitivně pro GUHA procedury realizující speciální „verze" metody GUHA. V některých procedurách (např. COLLAPS) je tištěna informace o redukci počtu otázek využitím podobných pravidel; četné příklady ukazují, že při větších rozsazích množiny generovaných otázek jde o redukci více než stonásobnou.

2.10 Zde naznačená implikační (IMPL) i asociační (ASSOC) procedura je realizována v systému GUHA programů s pracovním názvem GUHA 79.1. Tento systém obsahuje ještě další podobné procedury (COLLAPS, CORREL) a obslužné programy. Je koncipován jak pro samostatnou práci, tak pro přímou návaznost na softwarové systémy obecného charakteru pro analýzu dat (např. BMDP, viz [1]). Různé verze tohoto systému jsou nyní implementovány v ČSSR asi na dvaceti počítačích typů IBM 370/135, IBM 370/148, EC 1040, EC 1032 a ICL-System 4 a jsou používány řadou pracovišť, ČSAV počínaje a Podnikem automatizace řízení OKR konče. Systém obsahuje řadu zobecnění, o kterých bude řeč v odstavci IV.

Na jeho rozvoji se podílí řada programátorů: RNDr. J. Rauch, D. Pokorný, RNDr. Petr Kůrka, CSc., RNDr. A. Sochorová, ing. B. Louvar, dr. P. Jirků, J. Hlavešová, RNDr. M. Liška, RNDr. J. Vosáhlo, M. Nedvěd.

III. Matematické otázky

Tuto kapitolu může čtenář zajímající se pouze o základní informaci o metodě GUHA při prvním čtení vynechat. Tyto věci mají však význam pro závěrečnou diskusi a zhodnocení přínosu metody.

Lehce načrtnutý obraz v přechozí části článku vzbuzuje mnoho matematických (logických) otázek. Jde zejména o otázky týkající se modifikace obvyklého predikátového počtu ve dvou směrech: použití zobecněných kvantifikátorů, např. $\Rightarrow$, $\Rightarrow_p$, $\sim^\circ$, $\sim_\alpha$, a omezení se na konečné (logické) modely – struktury. Tato modifikace přináší mnoho zajímavých a někdy i překvapivých výsledků. Logika tohoto typu nebyla dosud příliš zkoumána, zkoumaly se sice zobecněné kvantifikátory, ale jiného typu (např. $\exists^\omega$-existuje nekonečně mnoho).

3.1 První z matematických otázek je otázka *definovatelnosti* kvantifikátorů. Zjistilo se, že mnoho zobecněných kvantifikátorů používaných v metodě GUHA není klasicky definovatelných, neboť k sentencím s jejich pomocí vytvořeným neexistují obecně sentence klasického predikátového počtu s nimi ekvivalentní. Přitom jde v případě logiky, která nás zajímá z hlediska metody GUHA, o definovatelnost v predikátovém počtu

s konečnými modely (data jsou vždy konečná). Zjišťování definovatelnosti není ani v tomto případě triviální otázkou. Jedním z nejjednodušších příkladů klasicky nedefinovatelného kvantifikátoru je $\sim^{\circ}$ uvedený v 2.5. Tyto otázky jsou řešeny např. v [7].

3.2 Další otázkou je problém *rozhodnutelnosti*. Logický kalkul můžeme zhruba chápat jako objekt, který sestáva jednak z formulí, sentencí a pravidel na jejich vytváření, jednak z určité třídy struktur, ve kterých mají být tyto sentence a formule obecně interpretovány, tedy například sentencím mají být v závislosti na dané struktuře z dané třídy struktur připisovány hodnoty „pravda" nebo „lež". Sentence pravdivá v každé struktuře (modelu) z dané příslušné třídy struktur se nazývá tautologií. Otázka rozhodnutelnosti může být nyní vyslovena takto: jde o to, zda pro daný kalkul je množina sentencí rozpoznatelná nějakou algoritmickou procedurou, zda tedy existuje algoritmus, který by nám pro každou sentenci kalkulu řekl, zda je či není tautologií.

Je možné říci, že predikátový počet s konečnými modely má v této otázce zajímavé chování. Platí totiž toto tvrzení (vždy uvažujeme predikátový počet s konečnými modely):

(1) Existuje predikátový počet s identitou, konečně mnoha predikáty a kvantifikátory $\forall$ a $\exists$, který *není rozhodnutelný*.

(2) Každý predikátový počet s konečně mnoha unárními predikáty *je rozhodnutelný*.

(3) Každý predikátový počet bez identity s konečně unárními predikáty a konečně mnoha (zobecněnými) kvantifikátory *je rozhodnutelný*.

(4) Existuje predikátový počet s pouze unárními predikáty splňující libovolnou z následujících podmínek:

(i) jediný predikát a nekonečně mnoho kvantifikátorů,

(ii) nekonečně mnoho predikátů a dva kvantifikátory,

(iii) identita, konečně mnoho predikátů a kvantifikátorů,

který *není rozhodnutelný*.

Je tedy zřejmé, že se v této oblasti pohybujeme na hranicích rozhodnutelného a nerozhodnutelného.

Je možné také uvažovat i o rozhodnutelnosti pouze určité podmnožiny sentencí daného tvaru. V této oblasti jsou některé pozoruhodné výsledky (viz [7]). Zajímavé je, že důkazy jsou poměrně komplikované a vyžádaly si použití i na první pohled dosti odlehlých výsledků, např. Scottova výsledku týkajícího se reprezentace kvalitativní pravděpodobnosti na konečných polích jevů [26].

3.3 Další okruh otázek týkající se logiky konečných modelů je spjat s *teorií složitosti*, a tedy s matematickými základy „computer science".

Velmi snadno se ukáže, že pro velmi jednoduché množiny sentencí RQ může být nalezení řešení $X(\boldsymbol{M})$ otázkou procedur, které jsou velmi složité. Příklad: uvažujme n predikátů $V_1, \ldots, V_n$ (unárních). Nechť pro dané $e = \{i_1, \ldots, i_k\} \subseteq \{1, \ldots, n\}$ říká sentence φ_e toto: každý objekt nemá alespoň jednu z vlastností $V_{i_1}, \ldots, V_{i_k}$. Zřejmě lze použít pravidla dedukce:

$$I = \left\{ \frac{\varphi_e}{\varphi_{e'}} : e \subseteq e' \right\}.$$

Položíme-li otázky $RQ = \{\varphi_e : e \subseteq \{1, \ldots, n\}\}$, pak minimální řešení $X(\boldsymbol{M})$ obsahuje právě sentence φ_e pravdivé v $\boldsymbol{M}$ a takové, že žádné $\varphi_{e-\{i\}}$ není pravdivé v $\boldsymbol{M}$. Je možno relativně snadno ukázat, že algoritmus, který by „tiskl" $X(\boldsymbol{M})$, musí být nejméně exponenciálně složitý (v počtu kroků jako funkci počtu sloupců v matici $\boldsymbol{M}$). Exponenciální procedury se považují za prakticky nerealizovatelné. Zde nám ovšem pomůže jisté předem dané rozumné omezení na rozsah dat — počet sloupců.

V teorii složitosti hrají důležitou roli třídy jazyků rozpoznatelných Turingovým strojem v polynomiálním čase (tj. čase počítaném v počtu kroků hlavy, který polynomiálně závisí na délce vstupního slova). Tato třída se označuje P. Problémy, které lze chápat jako problém rozpoznání určitého jazyka z třídy P, lze považovat za strojově řešitelné. Podobně NP je třída jazyků rozpoznatelných v polynomiáním čase na nedeterministickém Turingově stroji.

Řekneme, že NP jazyk (problém) je univerzální NP jazyk (problém), dá-li se na něj libovolný NP jazyk (problém) redukovat polynomiálně složitě. Pudlák [20] ukázal, že dva problémy spjaté s metodou GUHA jsou NP univerzální:

(1) Problém pravdivé sentence: Dána matice nul a jedniček $\boldsymbol{M}$ rozměrů $m \times n$ a číslo $k \in \mathcal{N}$. Problém záleží v určení, zda existuje sentence φ_e pravdivá v $\boldsymbol{M}$, taková že e má nejvýše k prvků.

(2) Problém postačující množiny sentencí: Dána konečná množina *Sent* sentencí, dedukční pravidlo I a číslo $k \in \mathcal{N}$. Úkolem je zjistit, zda existuje $X \subseteq Sent$, card $(X) \leqq$ $\leqq k$, takové, že $I(X) \supseteq Sent$.

Tento výsledek ukazuje na praktickou nerealizovatelnost předběžného algoritmu, který by před vlastním výpočtem zjistil, zda data obsahují „něco zajímavého" (1), nebo na jisté potíže s dedukcí (2).

Další výsledky z této oblasti obsahuje práce Pudláka a Springsteela [21]; uveďme si některé z nich:

(3) Pro libovolná data $\boldsymbol{M}$ (rozměrů $m \times n$) je možno rozhodnout v *polynomiálním* čase, zda existuje elementární disjunkce ψ délky n pravdivá v datech $\boldsymbol{M}$ (přesněji $\forall\psi$ pravdivé v $\boldsymbol{M}$).

(4) Pro libovolná data $\boldsymbol{M}$ (rozměru $m \times n$) a dané přirozené $k \leqq n$ je možné rozhodnout v čase 2^{log^2}, zda existuje elementární disjunkce pravdivá v $\boldsymbol{M}$ délky nejvýše k.

V (3) je proměnnou polynomu $max(m, n)$. Rozhodnutelnost v čase 2^{log^2} znamená, že existuje deterministický Turingův stroj rozhodující problém v čase $2^{c\ log^2 h}$, kde $c \geqq 1$, *log* má základ 2 a v našem případě $h = mn$.

Případ (4) tedy leží mezi P a NP složitostí (přesněji leží pod exponenciální složitostí 2^h, která se rovněž považuje za příliš velkou pro strojovou realizaci).

Citovaná práce obsahuje mnoho dalších výsledků týkajících se kvantifikátorů $\Rightarrow_p$, $\sim^\circ$, dat s neúplnou informací atd. Ale již z uvedených ukázek výsledků je zřejmé, že *zkoumané procedury se pohybují na hranici mezi věcmi počítačově realizovatelnými a nerealizovatelnými*; jejich zkoumání tedy může i přispět ke zmapování této hranice.

IV. Kontext analýzy dat a další zobecnění

4.1 Zaměřme se nejprve blíže ke kontextu jiných procedur pro analýzu empirických dat v observačních či exploračních studiích. Těchto procedur, zčásti založených na statistické teorii nebo používajících některé statistické testy a rozhodovací pravidla, je nepřeberná řada. Metoda GUHA obecně popisuje jisté rysy některých z těchto metod. Procedury, které jsou jednotlivými instancemi metody GUHA, jsou případy procedur pro analýzu dat. Obecný popis by měl být užitečný právě v procedurách, které používají jistých vlastností rozhodovacích kroků na datech (vlastností ne pravděpodobnostních, ale vlastností deterministického charakteru) k urychlení vlastního výpočtu (toto urychlení často znamená i principiální uskutečnitelnost). Příkladem buďtež procedury pro hledání nejlepší množiny regresorů v mnohorozměrné lineární regresi. Zde se snažíme pomocí lineární kombinace některých veličin, např. $V_1, \ldots, V_{30}$, vyjádřit co nejlépe veličinu V_{31} (na základě dat). Zajímá nás ovšem, zda pro některé $V_{i_1}, \ldots, V_{i_k}$ $(k < 30)$ nedostaneme již dostatečně dobré vyjadření V_{31}. To ovšem znamená hledání, pro dané k, množiny $V_{i_1}, \ldots, V_{i_k}$, která nejlépe vystihuje V_{31}. Mírou je reziduální součet čtverců $RSS(i_1, \ldots, i_k)$. Důležitý fakt je, že deterministicky $RSS(j_1, \ldots, j_{k+1}) \leqq RSS(i_1, \ldots, i_k)$ pro libovolné indexy $\{i_1, \ldots, i_k\} \subseteq \{j_1, \ldots, j_{k+1}\}$. Tato skutečnost je pak používána k přeskakování segmentů zcela podobně jako v procedurách IMPL, ASSOC či COLLAPS. Takto je konstruována již efektivně pracující procedura (viz LaMotteho procedura SELECT [15]). Tím netvrdíme, že tato procedura je odvozena od metody GUHA, ale argumentujeme pro to, že dedukce na datové úrovni je často užitečným prostředkem a patrně by byla ještě častěji užívána při přesném formálním popisu procedur.

4.2 Pro účely analýzy dat je možné a nutné v tomto textu zmíněné logické prostředky dále zobecňovat. Jedním ze směrů takového zobecňování je připouštění zpracovávaných struktur s obecnějšími hodnotami, než je $\{0, 1\}$, např. $\mathscr{V}_i = \{0, 1, \ldots, k\}$, nebo $\mathscr{V}_i = \mathscr{R}$, tj. připouštění dat „nominálního", resp. „reálného" typu. První typ spolu s praktickými návrhy algoritmů i obecnou formalizací byl uvažován P. Hájkem [3]. V poněkud jiném směru se jím zabýval D. Pokorný, který navrhl proceduru COLLAPS (viz [18], [19] a [13]) pro hledání zdrojů závislosti v dvourozměrných frekvenčních tabulkách, tj. v souborech dat tvaru $\boldsymbol{M} = \langle M, \|V_1\|, \|V_2\| \rangle$, kde $\mathscr{V}_1 = \{1, \ldots, r\}$ a $\mathscr{V}_2 = \{1, \ldots, c\}$ (a je předpokládán jistý mechanismus pravděpodobnostního vzniku dat). Tato procedura by byla bez využití dedukce zcela nerealizovatelná. Při zpracování dat dosud zmíněnými procedurami metody GUHA jde zpravidla ze statistického hlediska o zpracování mnohorozměrných frekvenčních (kontingenčních) tabulek a jednotlivé kvantifikátory jsou inspirovány statistickými testy. V dosavadních procedurách se používají kvantifikátory binární (tj. spojující dvě otevřené formule); ve statistice lze nalézt rozumnou inspiraci i pro ternární, kvaternární atd. kvantifikátory postihující vícečetné vztahy.

Zobecněním na data reálného typu se zabýval Havránek [11], [14]. Navrhované procedury jsou opřeny o pořadové statistiky, které připouštějí určitou dedukci (z hodnot statistiky na jistých datech lze usuzovat na její hodnotu na „podobných" datech). Výsledné procedury jsou pak ryze nestatistického charakteru z hlediska požadavků simultánní statistické inference.

Dalším (nikoliv po stránce časové) zobecněním je zobecnění na data obsahující hodnotu „nevím, nezměřeno atp." (symbol $\times$). Toto zobecnění provedli Hájek, Bendová a Renc [3]. Použili hodnot $\{0, \times, 1\}$ a Kleeneův-Körnerův tříhodnotový kalkul. Principy uplatněné ve zmíněné práci lze rozšířit i na složitější data. Praktické uplatnění je spojeno s důkazy „monotonie" určitých statistik vůči změnám v datech, které mohou být z obecně matematického hlediska zcela nezajímavé, zato však dosti komplikované [22].

Poslední cesta rozvoje, nastoupená teprve nedávno, se týká obohacení našeho přístupu k analýze dat o některé prostředky z oblasti „umělého intelektu"; konkrétně jde o využití Lenatových ([16], [17]) myšlenek, na jejichž základě konstruoval svůj systém AM pro „objevování" v matematice, pro konstrukci automaticky se řídícího systému analýzy dat založeného jak na dosavadních GUHA-procedurách, tak na některých dalších procedurách analýzy dat.

4.3 Dosavadní procedury metody GUHA pracují hlavně na speciálním typu dat – dvouhodnotových (či někdy vícehodnotových) mnohorozměrných frekvenčních tabulkách. Pro takové tabulky jsou známy teoreticky prozkoumané statistické procedury, které je mohou analyzovat. Tyto procedury mají však dvě vady: Pro větší dimenze tabulek ($n = 6, 7, 8$) je množství otázek, které je třeba prozkoumat (hypotéz, které by bylo možné testovat) opět značně velké a navíc pro dimenze větší než 7 jsou tyto postupy zpravidla počítačově nerealizovatelné a jejich asymptotická teorie zpravidla vzhledem k počtu pozorovaných objektů neaplikovatelná. Obvykle se při analýze takových tabulek (vznikajících v praxi velmi často na základě lékařských nebo sociologických výzkumů) využívá procedur vytvářejících podtabulky, např. podtabulky tvořené objekty splňujícími určité derivované vlastnosti a obsahující údaje o některých dalších vlastnostech; na takovã data se aplikují klasické statistické testy (bez ohledu na problematiku simultánní statistické inference). Procedury metody GUHA nedělají pro tento typ dat vlastně nic jiného, než že systematickým a optimálním způsobem vytvářejí a vyhodnocují takové podtabulky určitého typu. Je možno jen opakovat, že zde je cesta dalšího rozvoje opřena o inspiraci ke zkoumání složitější vztahů (viz 4.2). Důležité je, že stejně jako pro jiná zobecnění je zde příslušný logický jazyk již vytvořen [7].

Zbývá se zmínit o otázce celkové spolehlivosti výsledků získaných procedurami metody GUHA. Otázky, resp. kvantifikátory, jsou často tvořeny na základě statistických testů a u těch jsme zvyklí znát spolehlivost jejich výsledků. Zde lze obecně říci, že stejně jako i u mnoha dalších procedur pro analýzu dat není taková spolehlivost cílem; parametry použitých testů jsou tu spíše chápány jako měřítko pro srovnávání jednotlivých odpovědí, nikoliv jako vyjadřování jejich absolutní spolehlivosti. Cílem je *nalézt pokud možno všechny hypotézy, které by mohly zodpovědět náš globální cíl výzkumu*. Jde o *hypotézy*, které jsou sice na základě dat do jisté míry racionálně zdůvodnitelné, ale je nutné s nimi dále jako z hypotézami zacházet – prověřovat jejich konzistenci se známými skutečnostmi daného oboru a zkoumat jejich konzistenci s nově získanými empirickým daty. Na tomto celkovém obrazu nic nemění některé Havránkovy výsledky týkající se dílčí spolehlivosti některých procedur.

V. Aplikace

Aplikace metody GUHA (původních implikačních a asociačních procedur pro $\{0, 1\}$ hodnotová data a s kvantifikátory, které jsou různými zobecněními $\Rightarrow$, $\Rightarrow_p$, $\sim_\alpha$) byly nepříliš četné alespoň z hlediska publikovaných prací, ve kterých byla použita: v letech 1967–1975 jde řádově o patnáct prací. Je ovšem nutno si uvědomit dvě skutečnosti:

(1) Původní programy byly vytvořeny pro počítače MINSK 22, které v uvedené době byly postupně vyřazovány z provozu.

(2) Metoda GUHA je metodou na generování hypotéz – a hypotézy se vlastně nepublikují.

Skutečností je, že metoda byla rutinně využívána na několika pracovištích (hlavně pro aplikovaný sociologický výzkum). Aplikace nebyly zpravidla příliš rafinované a nevyužívaly všech možností metody. Obor aplikací byl omezen připuštěným typem zpracovávaných dat – šlo zpravidla o zpracovávání dotazníku s odpověďmi ano–ne a podobná data, hlavně v oblasti lékařské a sociologické, ale nechybí ani aplikace v technické diagnostice příčin poruch apod. Čtenáře by mohly zajímat aplikace popisované v článcích Z. Rence [23], [24], [25].

Nové aplikace počínají rokem 1977, kdy byla dokončena první verze nynějšího systému programů. První aplikací provedenou pomocí těchto programů je opět aplikace sociologická [12]. Zmíníme se nyní konkrétněji o dvou aplikacích (jedné z oblasti sociologické a jedné medicínské).

5.1 První aplikace, o které bude řeč, je pokračováním [12]. První krok popsaný v [12] byl pilotní studií, která měla ověřit metodiku výzkumu včetně například dotazníku. Proto byl použit poměrně malý počet dotazovaných osob (126). Vlastní studie pak použila reprezentativní výběr 1200 respondentů. Předmětem zkoumání byl vztah sociálních deskriptorů (vzdělání, funkce, obor zaměstnání atd.) k provozování různých kulturních aktivit (počínaje návštěvou kina a konče hrou v amatérském symfonickém orchestru). Šlo po úpravách (vypuštění veličin s velmi málo frekventovanými odpověďmi atp.) o něco přes sto vlastností. Byla použita jak asociační, tak implikační procedura. Důraz byl z hlediska zadavatele kladen na vyhledání všech možných vztahů a to vedlo k použití převážně asociační procedury s liberálně volenými kritérii výběru „významných" vztahů. Použil se tedy například kvantifikátor založený na chíkvadrát testu s 5% hladinou významnosti. Obdržené výsledky byly tvaru

muž & VŠ vzdělání & město $\sim$ pěstování umělecké fotografie.

Takových výsledků byla ovšem velká řada. Ukázaly na značné množství souvislostí, ale zejména pomohly nalézt určité typické kulturní aktivity pro různé skupiny podle sociálních deskriptorů.

Za zmínku snad stojí, že program REPORT ze systému GUHA 79.1 umožnuje tisknout výsledky doslova ve výše uvedené verbální formě (samozřejmě bez diakritických znamének). Výsledky v této formě sice neumožňují srovnávat různé vytištěné sentence (hypotézy), ale jsou velice přehledné pro první studium výsledků. Program REPORT může ovšem doplnit tisk výsledků o frekvence a, b, c, d a další statistiky vždy pro každou tištěnou sentenci. To je vhodné pro podrobnější zkoumání výsledků. Velmi zajímavé se

ukázaly výsledky v oblasti příčin, pro které považují respondenti subjektivně své kulturní „využití" za nedostatečné. Velmi symptomatický byl vztah kombinací pohlaví a vzdělání k uváděným příčinám, zejména v souvislostí s uvedenou překážkou „dítě": pro čtenáře Pokroků uvádíme jim jistě známý fakt, že pro vysokoškolsky vzdělané osoby nehraje v postoji k této otázce roli pohlaví respondenta.

Je nutné se zmínit o tom, že k analýze vztahů, zejména po upozornění procedurou ASSOC na určité souvislosti, byly použity další metody analýzy dat, které pak tyto souvislosti podrobněji prozkoumaly. Tato věc je velice podstatná – *procedury metody GUHA nejsou samospasitelnou metodou analýzy dat a je vhodné je kombinovat s jinými metodami.*

Výzkum byl prováděn Ústavem pro výzkum kultury, Praha.

5.2 Dalším příkladem použití je zpracování rozsáhlého souboru dat prováděné ve spolupráci mezi Angiologickou laboratoří fakulty všeobecného lékařství UK a Ústavem výpočetní techniky ČVUT. Jde o údaje o zhruba 2000 mužích, na kterých byl sledován jejich zdravotní stav před pěti léty a případy výskytu ischemické choroby srdeční v následujícím pětiletém intervalu. Zde bylo nutné hledat například kombinace údajů o zdravotním stavu (ve formě elementárních konjunkcí), které „vylučují" výskyt ischemické choroby srdeční. Byly hledány sentence tvaru

„elementární konjunkce vlastností vyjadřující zdravotní stav"
$\Rightarrow_{0.99}$ ne výskyt ischemické choroby.

Takové sentence pravdivé v datech byly skutečně nalezeny a přinesly některé nové aspekty a pohledy na výskyt ischemické choroby. Byla tedy použita procedura IMPL. Tato procedura byla rovněž použita pro vyhledávání okolností, které vysoce zvyšují riziko vzniku choroby nad obvyklou populační mez; zde byl použit kvantifikátor $\Rightarrow_{0.1}$ tj. „skoro" implikace s nezvykle nízkým $p = 0.1$. Je však nutné si přitom uvědomit, že nalezení kombinace zvyšující riziko nad např. 20% (tj. 0.2) je za současného stavu znalosti úspěchem. Výsledky byly opět konfrontovány s dalšími metodami analýzy dat (diskriminační analýza, logistický odhad rizika). Uvedené metody se velice vhodně doplňují, zejména vzhledem k „smíšenému" charakteru dat.

Na tomto místě považuji za nutné vložit za svou osobu tuto poznámku: K popisu aplikací jsem byl po velkém zdráhání donucen recenzentem a redakcí. Moje váhání nebylo způsobeno tím, že by aplikace nebyly úspěšné, ale tím, že v časopiseckém článku, který navíc má být trochu i o jiných věcech, působí obvykle primitivně anebo nesrozumitelně. Věci, jejichž zajímavé a mnohostranné aspekty jsou obvykle vidět pouze z obsáhlých výzkumných zpráv s řadou statistických tabulek, je riskantní jen popisovat. O to horší je to u metod, kdy jde o objevování nových pohledů na určitou výzkumnou problematiku a kdy každý pohled, je-li vytržen ze souvislosti, obvykle působí podivně.

5.3 Co brání dalšímu rozšíření metody GUHA a jejím aplikacím v oblastech, kde by mohla účinně přispět k řešení výzkumných problémů? Patrně tři okolnosti:

(i) Většina literatury o ní je nematematikovi a někdy i matematikovi nikoli přímo potřebné specializace víceméně nesrozumitelná.*)

*) Poznámka redakce: Z podobných důvodů byla k otištění přijata teprve čtvrtá verze vyžádaného článku. Obě strany se dohodly na tom, že každá z nich upřímně oceňuje trpělivost druhé.

(ii) Zatím chyběly přístupné programy a napsat je podle algoritmů publikovaných v literatuře je dosti obtížné. Programy uvedené do chodu v letech 1977 až 1979 nejsou zdaleka ideální; jejich základní postup je překvapivě rychlý a výkonný, ale zbývá zde mnoho vykonat v otázce flexibilních a jednoduchých vstupů a výstupů. Tato otázka je pro širší používání nejen programů GUHA zcela vitální. V této oblasti se nyní velice intenzívně pracuje a systém dokončený ke konci roku 1980 je v tomto směru velkým pokrokem.

(iii) Když někdo použije chybně nějakou známou matematickou metodu (např. regresní analýzu), neprohlásí zpravidla, že tato metoda je hloupost a k ničemu se nehodí. U metod, které nejsou u odborné veřejnosti dostatečně zavedeny, je tomu zpravidla naopak.

V. Závěr

Jak můžeme nyní hodnotit patnáct let rozvoje metody GUHA? Viděli jsme, že otevřela dosti širokou matematickou problematiku dotýkající se matematické logiky, matematických základů „computer science" i matematické statistiky. Zpravidla šlo pod vlivem praktických potřeb o obrácení pozornosti k matematickým objektům, které nejsou příliš složité, s kterými se často setkáváme, ale které nebyly dosud zkoumány a dokonce nám velmi často chybějí i metody pro jejich matematické zkoumání. Příkladem je logika zobecněných kvantifikátorů na konečných modelech nebo výpočtová složitost a jiné nepravděpodobnostní vlastnosti statistik. Matematické otázky takto vzniklé jsou jak ryze teoretického charakteru, tak i praktické, jejichž zodpovědění si vyžadovala konstrukce konkrétních algoritmů a programů. Většina těchto otázek byla úspěšně řešena.

Z hlediska perspektiv dalšího rozvoje můžeme být rovněž spokojeni. Zdá se, že se podařilo otevřít velmi perspektivní cestu pro matematický výzkum, a to jak teoretický, tak i aplikační. Okruh matematiků, kteří se uvedenou problematikou zabývají, rovněž roste (60 prací dvanácti autorů v letech 1966–1980, viz [8] s téměř úplnou bibliografií do roku 1978).

Z hlediska aplikací je situace složitější. Obecně lze říci, že u netriviálních metod, které vzniknou na základě impulsů z praxe, je často velmi dlouhá cesta k jejich dokončení ve formě vhodné pro aplikace, a tedy k zpětnému uplatnění v praxi. Původní problém, pro který byly konstruovány, je již zcela neaktuální v době, kdy jsou schopny úspěšného použití. Uplatnění v praxi závisí na „dotažení" metody včetně popularizace a vhodných programů (to neplatí jenom o metodě GUHA, ale vzpomeňme si například na celou řadu metod mnohorozměrné statistiky). Práce vložená do složitějších metod je vždy jen pomalu návratná; nelze čekat bleskový úspěch v aplikacích už jen proto, že v experimentálních vědách je mnohdy cyklus práce od nápadu k publikaci mnohem pomalejší než v matematice, kdy závisí víceméně jen na schopnostech matematikových.

Pro skutečnou aplikovatelnost metody v širokém měřítku zůstávají její autoři ještě leccos dlužni, i když vzniklo několik vysvětlujících článků a je v tisku česká monografie, popisující mj. systém programů a podávající velmi široce koncipovaný komentář k jejich používání. Přiznejme si, že matematici se raději zabývají matematickými problémy než

obtížným vysvětlováním, programováním a „nečistými" konkrétními aplikacemi. Je to do značné míry způsobeno i hodnocením těchto činností z hlediska matematické obce. Tím zapadá problematika aplikací metody GUHA do živých souvislostí aplikací matematiky a jejich hodnocení vůbec.

Dovolím si přidat ještě subjektivní závěr týkající se mé, a nejen mé, víry v principiální aplikovatelnost metod formování hypotéz, metod automatické analýzy dat a speciálně metody GUHA. Domnívám se, že je to jedna z plodných cest pro překonání našeho tápání při zpracování rozsáhlých dat vznikajících nyní v mnoha observačních studiích (v lékařství, sociologii, ekologii a celé řadě i technických oborů; též díky automatizaci sběru dat), které při obvyklém tradičním zpracování nás hrozí zcela zavalit.

Literatura

Poznámka: úplnou bibliografii prací o metodě GUHA lze získat u autora tohoto článku.

[1] W. J. Dixon, M. B. Brown (eds.): *BMDP — Biomedical Computer Programs.* University of California Press, Los Angeles 1977.

[2] P. Hájek: *Obecné pojetí metody GUHA.* Kybernetika *4* (1968), 505—515.

[3] P. Hájek: *Automatic listing of important observational statements.* Kybernetika *9*(1973), 187—205, 251—271 a *10* (1974), 95—124.

[4] P. Hájek, K. Bendová, Z. Renc: *The GUHA method and the three-valued logic.* Kybernetika *7* (1971), 431—435.

[5] P. Hájek, I. Havel, M. Chytil: *Metoda GUHA automatického vyhledávání hypotéz I, II.* Kybernetika *2* (1966), 31—47 a *3* (1967), 430—437.

[6] P. Hájek, I. Havel, M. Chytil: *The GUHA method of automatic hypotheses determination.* Computing *1* (1966), 293—308.

[7] P. Hájek, T. Havránek: *Mechanizing hypothesis formation — mathematical foundations for a general theory.* Universitext, Springer Verlag, Heidelberg—Berlin—New York, 1978.

[8] P. Hájek, T. Havránek: *The GUHA method — its aims and techniques.* Int. J. Man-Machine Studies *10* (1978), 3—22.

[9] T. Havránek: *Statistical quantifiers in observational calculi.* Theory and decision *6* (1975), 213—230.

[10] T. Havránek: *On simultaneous inference in contingency tables.* Aplikace matematiky *23* (1978), 31—38.

[11] T. Havránek: *Enumeration calculi and rank methods.* Int. J. Man-Machine Studies *10* (1978), 59—66.

[12] T. Havránek, M. Chyba, D. Pokorný: *Processing sociological data by the GUHA method — an example.* Int. J. Man-Machine Studies *9* (1977), 47—58.

[13] T. Havránek, D. Pokorný: *GUHA-style processing of mixed data.* Int. J. Man-Machine Studies *10* (1978), 47—58.

[14] T. Havránek, J. Vosáhlo: *A GUHA procedure with correlational quantifiers.* Int. J. Man-Machine Studies *10* (1978), 67—74.

[15] R. R. Hocking: *The analysis and selection of variables in linear regression.* Biometrics *32* (1976), 1—49.

[16] D. B. Lenat: *Automated theory formation in mathematics.* Fifth IJCAI, Cambridge 1977, 833 — 842.

[17] D. B. Lenat: *The ubiquity of discovery.* Artificial Intelligence *9* (1977), 257—285.

[18] D. Pokorný: *The GUHA method and desk calculators.* Int. J. Man-Machine Studies *10* (1978), 75—86.

[19] D. POKORNÝ, T. HAVRÁNEK: *On some procedures for identifying sources of dependence in contingency tables.* COMPSTAT 1978, L. C. A. CORSTEN, J. HERMANS (eds.) Physica-Verlag, Wien 1978, 221–227.

[20] P. PUDLÁK: *Polynomially complete problems in the logic of discovery.* MFCS '75, J. BEČVÁR (ed.), Lecture Notes in Computer Science *32*, Springer-Verlag, Heidelberg–Berlin–New York 1975, 358–361.

[21] P. PUDLÁK, F. N. SPRINGSTEEL: *Complexity in mechanized hypothesis formation.* Theoretical Computer Science (v tisku).

[22] J. RAUCH: *Ein Beitrag zu der GUHA Methode in der dreiwertigen Logik.* Kybernetika *11* (1975), 101–113.

[23] Z. RENC: *Průzkum uchazečů o studium na matematicko-fyzikální fakultě UK.* Sociologický časopis *5* (1972), 527–540.

[24] Z. RENC: *On interpretation of GUHA results.* Int. J. Man-Machine Studies *10* (1978), 37–46.

[25] Z. RENC, K. KUBÁT, K. KOUŘIM: *An appliacation of the GUHA method in medicine.* Int. J. Man-Machine Studies *10* (1978), 29–36.

[26] D. S. SCOTT: *Measurement structures and linear inequalities.* J. Math. Psychology *1* (1964), 233–247.

Hausdorffův-Banachův-Tarského paradox*)

L. E. Dubins

Jeho znění je (přesné definice budou podány později): „Jestliže X a Y jsou dvě ohraničené podmnožiny v $\mathbf{R}^3$ s neprázdnými vnitřky – např. jablko a Měsíc –, pak je možno rozdělit X na konečný počet částí a přemístit je tak, že vytvoří Y".

Podstata paradoxu je obsažena v následujícím lemmatu, které tvoří jeho geometrickou část.

Lemma. *Nechť S je jednotková sféra v $\mathbf{R}^3$ (o rovnici $x^2 + y^2 + z^2 = 1$). Existují dvě rotace a, b grupy SO_3 o úhly $180°$, resp. $120°$ a rozklad (A,B,C,D) sféry S s tou vlastností, že D je spočetná množina a platí*

$$C = bB = b^2A, \quad A = a(B \cup C).$$

Jinak řečeno množina A je současně třetinou i polovinou množiny $S - D$.

*) *Le paradoxe de Hausdorff-Banach-Tarski* (Fragment d'un cours de L. E. DUBINS rédigé par M. EMERY), Gazette des mathématiciens, N° 12 (Août 1979), 71–76. Přeložil IVAN KOLÁŘ.