Aplikace matematiky

Zdeněk Renc

# Automatická binarisace veličin

*Aplikace matematiky*, Vol. 15 (1970), No. 2, 117–124

Persistent URL: http://dml.cz/dmlcz/103275

# AUTOMATICKÁ BINARISACE VELIČIN[1])

ZDENĚK RENC

(Došlo dne 11. října 1968)

## I

Experimentální pracovník libovolného oboru (na př. lékařského biologického, sociologického atd.) se patrně dosti často ocitne v následující situaci. Shromáždil experimentální materiál skládající se z nějakých objektů (na př. pacientů, obyvatel nějaké oblasti, fyzikálních měření) a různých dat, která o těchto objektech zjistil. Pro svou další práci potřebuje zjišťovat různé zákonitosti a souvislosti mezi těmito daty. Abychom tuto situaci mohli lépe popsat, předpokládejme, že máme dánu množinu $\boldsymbol{M}$ jakýchkoli $m$ objektů a že jsou dále dány binární vlastnosti $P_1, \ldots, P_n$, $A_1, \ldots, A_l$ a veličina $V$, při čemž pro každý objekt z naší množiny $\boldsymbol{M}$ je známo jednak zda splňuje po řadě vlastnosti $P_1, \ldots, A_l$, jednak jakou hodnotu mu připisuje veličina $V$. Předpokládejme dále, že $P_1, \ldots, P_n$ jsou zadány tak, že každé $x \in \boldsymbol{M}$ splňuje právě jednu z nich a že každá z vlastností $P_1, \ldots, P_n$ je splněna alespoň jedním objektem množiny $\boldsymbol{M}$. (Např. $P_1, \ldots, P_{n-1}$ mohou být různé vzájemně se vylučující druhy téže choroby, $P_n$ znamená „býti zdráv".) Máme zkoumat vztahy mezi $P_1, \ldots, P_n$ na jedné straně a $A_1, \ldots, A_l$ (což v našem příkladě mohou být nějaké další vlastnosti, které jsme u pacientů zjistili) a veličinou $V$ (např. věkem pacienta) na straně druhé. Tyto vztahy je možno zjišťovat na příklad pomocí metody GUHA, popsané v pracích [2], [3], [4], na které čtenáře odkazuji.

Abychom mohli metodu GUHA použít, musíme ovšem veličinu $V$ binarizovat nebo, chceme-li, nahradit ji systémem binárních veličin $V_1, \ldots, V_k$ (to znamená, že otázku „jaký je věk pacienta" potřebujeme nahradit systémem otázek „je věk pacienta v rozmezí $L_1$ až $L_2$ let" atd.) Aby byla binarisace veličiny $V$ rozumná, požadujeme, aby opět každý prvek $x \in \boldsymbol{M}$ splňoval právě jednu z vlastností $V_1, \ldots, V_k$ a aby každá z $V_1, \ldots, V_k$ byla splněna alespoň jedním objektem množiny $\boldsymbol{M}$ (to znamená, aby intervaly, na které jsme věkovou škálu rozdělili, se navzájem nepřekrývaly a aby dohromady pokryly celou tuto škálu).

Kromě této podmínky musíme mít ale na mysli ještě další požadavek, kterému můžeme říkat požadavek vhodnosti binarizace veličiny $V$ vzhledem k $P_1, \ldots, P_n$.

[1]) Tato práce byla referována na semináři aplikací matematické logiky na MFF KU a přednesena na XXIII. Fysiologických dnech, konaných v Železné Rudě.

Nechceme totiž binarizovat veličinu $V$ jakkoliv, nýbrž tak, aby tato binarizace byla co nejvýhodnější v tom smyslu, že nám pomůže rozdělit obor hodnot veličiny $V$ na částečné intervaly co nejlépe určující rozdělení prvků množiny $\boldsymbol{M}$ vzhledem k vlastnostem $P_1, \ldots, P_n$. Tuto vhodnost binarizace veličiny $V$ vzhledem k vlastnostem $P_1, \ldots, P_n$ budeme měřit způsobem, o kterém se čtenář může poučit v [1] a který nyní stručně popíšeme.

Představme si následující situaci. Máme dány binární vlastnosti $P_1, \ldots, P_n$ (s požadavky, o nichž jsme výše hovořili). Vybereme náhodně prvek $x \in \boldsymbol{M}$ a snažíme se co nejlépe „uhádnout" tu vlastnost z $P_1, \ldots, P_n$, kterou $x$ splňuje a to nejprve v první a potom v druhé z následujících dvou situací

1. Není nám nic známo o vlastnostech $V_1, \ldots, V_k$.
2. Je nám znám rozklad $V$ na systém binárních vlastností $V_1, \ldots, V_k$.

1. Nechť $a_{.1}, \ldots, a_{.n}$ jsou po řadě počty prvků z $\boldsymbol{M}$, splňujících $P_1, \ldots, P_n$, $a_{.\max} = \max_j a_{.j}$. Nejlepší odhad dostaneme, zvolíme-li tu vlastnost $P_j$, pro níž je $a_{.\max} = a_{.j}$. Předpokládáme-li, že pravděpodobnosti jevu „objekt $x \in \boldsymbol{M}$ splňuje vlastnost $P_j$" je rovna $P_{.j} = a_{.j}/m$ (to znamená, že tato pravděpodobnost je rovna relativní frekvenci $P_j$ v množině $\boldsymbol{M}$), můžeme říci, že nejlepší odhad dostaneme, zvolíme-li tu třídu $P_j$, pro níž je $p_{.\max} = p_{.j}$ (kde $p_{.\max} = \max_j p_{.j}$). Chyba tohoto odhadu je $1 - p_{.\max}$.

2. Nechť $a_{i1}, \ldots, a_{in}$ (pro $i = 1, \ldots, k$) jsou po řadě počty prvků z $M$, splňujících $V_i$ a současně $P_1, \ldots, P_n$, $a_{i\max} = \max_j a_{ij}$ pro $i = 1, \ldots, k$. Nechť $p_{i1}, \ldots, p_{in}$ $(i = 1, \ldots, k)$, $p_{i\max} = \max_j p_{ij}$ jsou odpovídající relativní frekvence (o nichž opět předpokládáme, že se rovnají pravděpodobnostem příslušných jevů). Nejlepší odhad dostaneme takto: je-li $x \in V_1$, splňuje tu vlastnost $P_j$, pro níž je $p_{1\max} = p_{1j}, \ldots$, je-li $x \in V_k$, splňuje tu vlastnost $p_j$, pro níž je $p_{k\max} = p_{kj}$. Chyba tohoto odhadu je $1 - \sum_{i=1}^{k} p_{i\max}$. Zřejmě chyba v situaci 2. nemůže být větší než v situaci 1. Označme

$$\lambda = \frac{1 - p_{.\max} - \left(1 - \sum_{i=1}^{k} p_{i\max}\right)}{1 - p_{.\max}} = \frac{\sum_{i=1}^{k} p_{i\max} - p_{.\max}}{1 - p_{.\max}}$$

$\lambda$ je tedy relativní pokles chyby našeho odhadu při přechodu od situace 1. k situaci 2. Čím větší je hodnota čísla $\lambda$, tím lépe nám znalost rozkladu veličiny $V$ na systém binárních vlastností $V_1, \ldots, V_k$ pomáhá určit rozdělení prvků množiny $\boldsymbol{M}$ podle vlastností $P_1, \ldots, P_n$.

Úkolem, který se budeme v této práci snažit vyřešit, bude (zhruba řečeno) nalézt pro dané vlastnosti $P_1, \ldots, P_n$ a danou veličinu $V$ nejvhodnější binarizaci této veličiny $V$ (ve smyslu naznačeném výše). V části II. zavedeme přesně pojmy, které jsme zde názorně vyložili a dokážeme ty jejich vlastnosti, které nám v části III. umožní popsat algoritmus pro práci samočinného počítače.[2])

[2]) Na potřebu podobných úvah upozornil již dříve ing. M. CHYTIL.

## II

**Definice.** *Modelem nazveme* $(n + 2)$*-tici* $\mathfrak{M} = \langle \boldsymbol{M}, P_1, \ldots, P_n, V \rangle$ *takovou, že*

(a) $\boldsymbol{M}$ *je konečná množina o* $m$ *prvcích*

(b) $P_1, \ldots, P_n$ *jsou unární predikáty na* $\boldsymbol{M}$ *takové, že každé* $x \in \boldsymbol{M}$ *splňuje právě jeden z* $P_1, \ldots, P_n$ *a že každý z* $P_1, \ldots, P_n$ *je splněn alespoň jedním* $y \in \boldsymbol{M}$

(c) $V$ *je reálná funkce na* $\boldsymbol{M}$.

Nechť $V$ nabývá na $\boldsymbol{M}$ $h$ $(h \geqq 3)$ různých hodnot. Uspořádejme je podle velikosti a označme postupně $1, \ldots, h$. Tomu odpovídá částečné uspořádání množiny $\boldsymbol{M}$ takové, že na prvních $m_1$ prvcích nabývá $V$ hodnoty 1, na dalších $m_2$ prvcích hodnoty 2, atd. $(m_1 + \ldots + m_h = m)$. Jestliže $V$ nabývá na $\boldsymbol{M}$ právě $h$ hodnot $1, \ldots, h$, řekneme, že $\mathfrak{M}$ je dobrý model. Předpokládáme v dalším, že model $\mathfrak{M}$ je dobrý.

**Definice.** *Řekneme, že* $\boldsymbol{D} = \{h_1, \ldots, h_k\}$ *je dělení intervalu* $\langle 1, h \rangle$ *stupně* $k$ $(2 \leqq k \leqq h)$, *jestliže* $h_1, \ldots, h_k$ *je vybraná posloupnost z posloupnosti* $1, \ldots, h$ *taková, že* $1 \leqq h_1 < h_2 < \ldots < h_k = h$.

**Definice.** *Rozkladem stupně* $k$ *k* $\boldsymbol{R}^k = \{I_1, \ldots, I_k\}$ *funkce* $V$ *na množině* $\boldsymbol{M}$, *indukovaným dělením* $\boldsymbol{D} = \{h_1, \ldots, h_k\}$ *rozumíme systém částečných intervalů* $I_1 = \langle 1, h_1 \rangle$, $I_2 = (h_1, h_2 \rangle, \ldots, I_k = (h_{k-1}, h_k \rangle$ *intervalu* $\langle 1, h \rangle$.

**Definice.** $V_1, \ldots, V_k$ *buďte unární predikáty, definované na množině* $\boldsymbol{M}$ *takto*: $V_i(x) \equiv V(x) \in I_i$, $i = 1, \ldots, k$.

K danému rozkladu veličiny $V$ (neboli k danému dělení $\boldsymbol{D}$) sestavme tabulku:

| | $P_1$ | $P_2$ | $\ldots$ | $P_n$ | |
|---|---|---|---|---|---|
| $V_1$ | $a_{11}$ | $a_{12}$ | $\ldots$ | $a_{1n}$ | $a_{1.}$ |
| $V_2$ | $a_{21}$ | $a_{22}$ | $\ldots$ | $a_{2n}$ | $a_{2.}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots\vdots\vdots$ | $\vdots$ | $\vdots$ |
| $V_k$ | $a_{k1}$ | $a_{k2}$ | $\ldots$ | $a_{kn}$ | $a_{k.}$ |
| | $a_{.1}$ | $a_{.2}$ | $\ldots$ | $a_{n.}$ | $m$ |

kde $a_{ij}$ je počet prvků množiny $\boldsymbol{M}$, které splňují současně $V_i$ a $P_j$ $(i = 1, \ldots, k, j = 1, \ldots, n)$, dále $a_{i.} = \sum_{j=1}^{n} a_{ij}$, $a_{.j} = \sum_{i=1}^{k} a_{ij}$. Položme $a_{i\max} = \operatorname{Max}_j a_{ij}$, $a_{.\max} = \operatorname{Max}_j a_{.j}$. Zřejmě platí $m = \sum_{j=1}^{n} a_{.j} = \sum_{i=1}^{k} a_{i.} = \sum_{i,j} a_{ij}$. Označme

$$\lambda_{\boldsymbol{D},\boldsymbol{M}} = \frac{\sum_{i=1}^{k} a_{i\max} - a_{.\max}}{m - a_{.\max}}$$

Označíme-li ještě $p_{i\max} = a_{i\max}/m$, $p_{.\max} = a_{.\max}/m$, je

$$\lambda_{\boldsymbol{D},\boldsymbol{M}} = \frac{m\left(\sum_{i=1}^{k} p_{i\max} - p_{.\max}\right)}{m(1 - p_{.\max})} = \frac{\sum_{i=1}^{k} p_{i\max} - p_{.\max}}{1 - p_{.\max}}$$

To znamená, že $\lambda_{\boldsymbol{D},\boldsymbol{M}}$ nezávisí na $m$ (t.j. počtu prvků množiny $\boldsymbol{M}$), neboli na faktických frekvencích v $\boldsymbol{M}$, nýbrž pouze na relativních frekvencích. Dále ho budeme tedy značit $\lambda_{\boldsymbol{D}}$.

Snadno se ověří následující vlastnosti čísla $\lambda_{\boldsymbol{D}}$ (srov. [1]).

(1) $\lambda_{\boldsymbol{D}}$ je definováno pro každé dělení $\boldsymbol{D}$.

(2) $0 \leqq \lambda_{\boldsymbol{D}} \leqq 1$ pro každé dělení $\boldsymbol{D}$.

(3) $\lambda_{\boldsymbol{D}} = 0$ právě když existuje $n_0$ přirozené tak, že $1 \leqq n_0 \leqq n$ a že pro všechna $i = 1, \ldots, k$ je $p_{in_0} = p_{i\max}$.

(4) $\lambda_{\boldsymbol{D}} = 1$ právě když pro každé $i = 1, \ldots, k$ existuje právě jedno $j$ tak, že $1 \leqq j \leqq n$ a že $p_{ij} \neq 0$.

(5) $\lambda_{\boldsymbol{D}}$ nezávisí na pořadí $V_1, \ldots, V_k$.

(6) jsou-li $V_1, \ldots, V_k$ a $P_1, \ldots, P_n$ statisticky nezávislé, je $\lambda_{\boldsymbol{D}} = 0$.

(3) říká, že $\lambda_{\boldsymbol{D}} = 0$, právě když znalost dělení $\boldsymbol{D}$ intervalu $\langle 1, h\rangle$ vůbec nepomáhá k lepšímu rozdělení prvků množiny $\boldsymbol{M}$ vzhledem k $P_1, \ldots, P_n$; naopak (4) znamená, že $\lambda_{\boldsymbol{D}} = 1$ právě když dělení $\boldsymbol{D}$ úplně určuje rozdělení prvků množiny $\boldsymbol{M}$ vzhledem k $P_1, \ldots, P_n$. Tvrzení (5) nám umožňuje převést každý model $\mathfrak{M}$ na dobrý model.

**Věta.** *Nechť $\boldsymbol{D}$ a $\overline{\boldsymbol{D}}$ jsou dvě dělení intervalu $\langle 1, h\rangle$ příslušného k veličině $V$ z $\mathfrak{M}$, nechť $\overline{\boldsymbol{D}}$ je zjemnění $\boldsymbol{D}$. Potom $\lambda_{\overline{\boldsymbol{D}}} \geqq \lambda_{\boldsymbol{D}}$.*

Důkaz. Označme $k$ resp. $\bar{k}$ stupeň dělení $\boldsymbol{D}$ resp. $\overline{\boldsymbol{D}}$. Jistě je $k \leqq \bar{k}$. Je-li $k = \bar{k}$, je $\boldsymbol{D} \equiv \overline{\boldsymbol{D}}$ a tedy $\lambda_{\boldsymbol{D}} = \lambda_{\overline{\boldsymbol{D}}}$. Nechť tedy $k < \bar{k}$. Zřejmě stačí důkaz provést pro případ $\bar{k} = k + 1$. Bez újmy na obecnosti můžeme předpokládat, že $\overline{\boldsymbol{D}}$ je tvořeno posloupností čísel $h_0, h_1, \ldots, h_k$, kde $1 \leqq h_0 < h_1 < h_2 < \ldots < h_k = h$, $h_0$ je některé z čísel $1, \ldots, h - 1$. Tabulka, odpovídající dělení $\overline{\boldsymbol{D}}$, je

| | $P_1$ | $P_2$ | ... | $P_n$ | |
|---|---|---|---|---|---|
| $V_0^1$ | $a_{01}^1$ | $a_{02}^1$ | ... | $a_{0n}^1$ | $a_{0.}^1$ |
| $V_0^2$ | $a_{01}^2$ | $a_{02}^2$ | ... | $a_{0n}^2$ | $a_{0.}^2$ |
| $V_2$ | $a_{21}$ | $a_{22}$ | ... | $a_{2n}$ | $a_{2.}$ |
| ⋮ | ⋮ | ⋮ | ⋮⋮⋮ | ⋮ | ⋮ |
| $V_k$ | $a_{k1}$ | $a_{k2}$ | ... | $a_{kn}$ | $a_{k.}$ |
| | $a_{.1}$ | $a_{.2}$ | ... | $a_{.n}$ | $m$ |

kde $a_{0j}^1 + a_{0j}^2 = a_{1j}$ pro $j = 1, \ldots, n$, $a_{0.}^1 + a_{0.}^2 = a_{1.}$. Tedy též $a_{0.}^1 = \sum_{j=1}^{n} a_{0j}^1$, $a_{0.}^2 = \sum_{j=1}^{n} a_{0j}^2$.

Položíme dále

$$a_{0\max}^1 = \operatorname{Max}_j a_{0j}^1, \quad a_{0\max}^2 = \operatorname{Max}_j a_{0j}^2,$$

tedy

$$\lambda_{\overline{D}} = \frac{a_{0\max}^1 + a_{0\max}^2 + \sum_{i=2}^{k} a_{i\max} - a_{.\max}}{m - a_{.\max}}$$

Jelikož

$$a_{1\max} = \operatorname{Max}_j a_{1j} = \operatorname{Max}_j (a_{0j}^1 + a_{0j}^2) \leqq \operatorname{Max}_j a_{0j}^1 + \operatorname{Max}_j a_{0j}^2 = a_{0\max}^1 + a_{0\max}^2$$

dostáváme ihned $\lambda_{\boldsymbol{D}} \leqq \lambda_{\overline{D}}$.

**Definice.** *Nechť $\boldsymbol{D}_1$ a $\boldsymbol{D}_2$ jsou dvě dělení intervalu $\langle 1, h\rangle$ příslušného k veličině $V$ z $\mathfrak{M}$. Řekneme, že $\boldsymbol{D}_1$ je vhodnější než $\boldsymbol{D}_2$ vzhledem k $P_1, \ldots, P_n$, jestliže $\lambda_{\boldsymbol{D}_1} \geqq \lambda_{\boldsymbol{D}_2}$.*

**Definice.** *Maximálním dělením $\boldsymbol{D}_{\max}$ intervalu $\langle 1, h\rangle$ nazveme dělení intervalu $\langle 1, h\rangle$, tvořené posloupností čísel $1, 2, \ldots, h$.*

Toto dělení je stupně $h$ a indukuje rozklad $\boldsymbol{R}^h$ veličiny $V$ na intervaly $I_1 = \langle 1, 1\rangle$,

$I_2 = (1, 2\rangle, \ldots, I_h = (h-1, h\rangle$ (tj. na „jednobodové intervaly"). Dělení $\boldsymbol{D}_{\max}$ je zjemněním každého dělení $\boldsymbol{D}$ intervalu $\langle 1, h\rangle$ a podle právě dokázané věty k němu příslušné $\lambda_{\max} \geqq \lambda_{\boldsymbol{D}}$ pro každé $\boldsymbol{D}$.

**Definice.** *Nechť* $\boldsymbol{D} = \{h_1, \ldots, h_k\}$ *je dělení intervalu* $\langle 1, h\rangle$, *které indukuje rozklady* $\boldsymbol{R}^k = \{I_1, \ldots, I_k\}$. *Normou dělení* $\boldsymbol{D}$ *nazveme číslo* $v(\boldsymbol{D}) = \operatorname{Min}(\bar{I}_i)$, *kde* $\bar{I}_i$ *je počet těch čísel z* $1, \ldots, h$, *která náleží do intervalu* $I_i$.

**Definice.** *Nechť* $\boldsymbol{D} = \{h_1, \ldots, h_k\}$ *je dělení intervalu* $\langle 1, h\rangle$, *které indukuje rozklad* $\boldsymbol{R}^k = \{I_1, \ldots, I_k\}$. *Normou dělení* $\boldsymbol{D}$ *vzhledem k množině* $\boldsymbol{M}$ *nazveme číslo* $v_{\boldsymbol{M}}(\boldsymbol{D}) = \operatorname{Min}_i(\bar{I}_i^{\boldsymbol{M}})$, *kde* $\bar{I}_i^{\boldsymbol{M}}$ *je počet prvků množiny* $\boldsymbol{M}$, *pro něž hodnota veličiny* $V$ *náleží do intervalu* $I_i$.

**Definice.** *Řekneme, že dělení* $\boldsymbol{D}$ *intervalu* $\langle 1, h\rangle$ *je prosté, jestliže pro každé dělení* $\overline{\boldsymbol{D}}$ *intervalu* $\langle 1, h\rangle$ *takové, že* $\boldsymbol{D}$ *je zjemněním* $\overline{\boldsymbol{D}}$, *platí* $\lambda_{\overline{\boldsymbol{D}}} < \lambda_{\boldsymbol{D}}$.

Naši úlohu můžeme nyní formulovat takto. Je-li $\mathfrak{M} = \langle \boldsymbol{M}, P_1, \ldots, P_n, V\rangle$ dobrý model, $k$, $d$ daná přirozená čísla ($2 \leqq k \leqq h$, $1 \leqq d \leqq h/2$), hledáme prosté dělení intervalu $\langle 1, h\rangle$ stupně nejvýše $k$, jehož norma $v$ je větší nebo rovna $d$, nejvýhodnější vzhledem k $P_1, \ldots, P_n$. Je-li takovýchto dělení více, uspořádáme je podle normy $v_{\boldsymbol{M}}$ vzhledem k množině $\boldsymbol{M}$.

## III

V této části popíšeme algoritmus řešení dané úlohy se zřetelem k jeho realisaci na samočinném počítači.

(A) Přípravná část.

1) Zjistit počet $h$ hodnot, které nabývá veličina $V$ na množině $\boldsymbol{M}$. Je-li $h = 1$, práce končí. Je-li $h = 2$, práce končí ($V$ je již binarizována).

2) Prověřit, zda některý z predikátů $P_1, \ldots, P_n$ není na $\boldsymbol{M}$ pravdivý nebo lživý. Pokud ano, vyřadit.

3) Ověřit, zda $d \leqq h/2$. Je-li $d > h/2$, práce končí.

4) Uspořádat funkční hodnoty veličiny $V$ podle velikosti, přiřadit jim čísla $1, \ldots, h$ a sestavit tabulku viz str. 123 (t.j. tabulku příslušnou k dělení $\boldsymbol{D}_{\max}$), kde $a_{ij}$ ($i = 1, \ldots, h$, $j = 1, \ldots, n$) je počet objektů $\boldsymbol{M}$, pro něž je veličina $V$ nabývá hodnoty $i$ a které splňují predikát $P_j$, $a_{i.}$, $a_{.j}$ příslušné součty, $a_{i\max}$, $a_{.\max}$ příslušná maxima.

5) Spočítat

$$\lambda_{\max} = \frac{\sum_{i=1}^{h} a_{i\max} - a_{.\max}}{m - a_{.\max}}$$

| | $P_1$ | ... | $P_n$ | |
|---|---|---|---|---|
| 1 | $a_{11}$ | ... | $a_{1n}$ | $a_{1.}$ |
| 2 | $a_{21}$ | ... | $a_{2n}$ | $a_{2.}$ |
| ⋮ | ⋮ | ⋮⋮⋮ | ⋮ | ⋮ |
| $h$ | $a_{h1}$ | ... | $a_{hn}$ | $a_{h.}$ |
| | $a_{.1}$ | ... | $a_{.n}$ | $m$ |

(B) Vlastní práce.

1) $\bar{k} = 2$. Generovat (v lexikografickém uspořádání) všechna dělení intervalu $\langle 1, h\rangle$ stupně 2 normy alespoň $d$. Pro každé dělení $\boldsymbol{D}$ zjistit $\lambda_{\boldsymbol{D}}$. Je-li větší nebo rovno dosud maximálnímu, zjistit ještě $v(\boldsymbol{D})$, $v_{\boldsymbol{M}}(\boldsymbol{D})$, jinak generovat další. Je-li $\lambda_{\boldsymbol{D}} = \lambda_{\max}$, vytisknout (spolu se zjištěnými údaji) a práce končí. Je-li $\lambda_{\boldsymbol{D}} < \lambda_{\max}$ přejít k dalšímu dělení. Pamatovat průběžně dělení s největším $\lambda$. Po skončení práce pro $\bar{k} = 2$ toto nejlepší dělení spolu se zjištěnými údaji vytisknout, je-li jich více, uspořádat nejprve podle normy $v$, potom podle normy $v_{\boldsymbol{M}}$. První z tištěných dělení ponechat v paměti.

2) $\bar{k} = \overline{k_0} + 1$. Ověřit $\bar{k} \leqq k$. Je-li $\bar{k} > k$, práce končí. Ověřit $d \leqq h/(\overline{k_0} + 1)$. Neplatí-li, práce končí. Generovat postupně (v lexikografickém uspořádání) dělení intervalu $\langle 1, h\rangle$ stupně $\bar{k}$ normy alespoň $d$. Pro každé zjistit $\lambda_{\boldsymbol{D}}$, je-li větší nebo rovno dosud maximálnímu a je-li prosté, zjistit ještě $v(\boldsymbol{D})$, $v_{\boldsymbol{M}}(\boldsymbol{D})$. Pamatovat průběžně prosté dělení s největším $\lambda$ (narazí-li se na dělení $\boldsymbol{D}$ takové, že $\lambda_{\boldsymbol{D}} = \lambda_{\max}$ tisknout a konec) a po skončení práce pro $\bar{k} = \bar{k}_0 + 1$ tisknout jako v 1).

3) $\bar{k} = k$. Ověřit $d \leqq h/k$. Neplatí-li, práce končí. Jinak generovat, pamatovat a tisknout jako ve 2). Konec.

*Literatura*


[1] *L. A. Goodman, W. H. Kruskal:* Measures of Association for Cross Clasifications, JASA, *49* (1954).

[2] *P. Hájek, I. Havel, M. Chytil:* GUHA-metoda systematického vyhledávání hypotéz, Kybernetika 2, 1966, 1, 31—47.

[3] *P. Hájek, I. Havel, M. Chytil:* GUHA-metoda systematického vyhledávání hypotéz II, Kybernetika 3, 1967, 5, 430—437.

[4] *P. Hájek:* Problém obecného pojetí metody GUHA, Kybernetika — v tisku.

Summary

AUTOMATIC BINARIZATION OF QUANTITIES

ZDENĚK RENC

When processing automatically biological (and other) experimental data the need often occurs to replace a certain measured quantity (e.g. pressure, electric potential, age etc.) by one or more properties (e.g. pressure heavy or light, temperature less than 30 °C, between 30 °C–60 °C, over 60 °C, etc.). As a property may be understood as a binary quantity, the problem of the replacement of a quantity by properties can be named the binarization problem. (This problem occurs, e.g., when applying the GUHA method of automatic hypotheses determination, see [2], [3], [4].) We ask which of possible binarizations is the best one (e.g. whether the binarization of temperature given above is better than the following one: less then 40 °C, between 40 °C–70 °C, over 70 °C). On the basis of the paper [1], it is possible to compare numerically which of two binarizations is more suitable with respect to one or more properties. In the present paper a method (called ABQ) is described for the automatic determination of the best binarization of a quantity with respect to given properties on the basis of given experimental material. When the research worker determines the maximal number of intervals that the quantity is to be divided into and the properties with respect to which the suitability is to be measured (and, of course, supplies the experimental data), the computer gives all the optimal binarizations. The computer counts also supplementary information concerning the binarizations obtained.

*Adresa autora:* RNDr. *Zdeněk Renc*, Matematicko-fyzikální fakulta KU, Sokolovská 83, Praha 8.