Časopis pro pěstování matematiky, roč. 102 (1977), Praha

SHLUKOVÁ ANALYSA

Adolf Filáček, Václav Koutník a Jiří Vondráček, Praha

(Došlo dne 1. června 1977)

Metody shlukové analysy (anglicky cluster analysis) se vyvinuly z potřeby analysovat a vhodně koncentrovat informaci obsaženou v mnohorozměrných údajích. Jednotlivé postupy řešící problém klasifikace údajů se objevují ve čtyřicátých a padesátých letech. Teprve v poslední době však dochází k pokusům vybudovat ucelenou teorii, která by umožňovala analysu vlastností jednotlivých postupů, jejich porovnání a posouzení, za jakých podmínek je možno je použít. Nicméně je shluková analysa stále spíše střechový název pro volný soubor heuristických postupů než ucelená disciplina aplikované matematiky.

Základní situaci při shlukové analyse můžeme popsat takto:

Je dáno N objektů. Na každém objektu je změřeno p charakteristik, takže získáváme N p-rozměrných vektorů $X_1, X_2, \ldots, X_N$. Můžeme ztotožnit pozorování a příslušné objekty, takže v dalším nazýváme vektory X_i objekty. Označme X množinu všech objektů, tj. $X = \{X_1, X_2, \ldots, X_N\}$. Úkolem shlukové analysy je seskupit objekty X_i do n shluků $S_1, S_2, \ldots, S_n$ (tj. množina $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$ tvoří rozklad množiny X) tak, aby si objekty patřící do téhož shluku byly v jistém smyslu podobné či blízké, kdežto od objektů patřících do různých shluků požadujeme, aby byly odlišné či vzdálené. Přitom obvykle chceme, aby počet shluků n byl podstatně menší než počet objektů N. Někdy je úloha shlukové analysy formulována obecněji v tom smyslu, že shluky nemusí být disjunktní, a cílem není rozklad, ale pokrytí množiny X jejími podmnožinami.

Podle cíle můžeme rozeznávat tři druhy úloh shlukové analysy:

1) Cílem je nalezení rozkladu $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$, kde počet shluků n není předem stanoven.

2) Cílem je nalezení rozkladu $\mathbf{S}$ při předem daném počtu shluků n.

3) Cílem je vytvořit tzv. hierarchický strom, tj. posloupnost rozkladů $\mathbf{S}^{(t)}$, $t = 1, 2, \ldots, K$, kde $\mathbf{S}^{(1)} = \{\{X_1\}, \{X_2\}, \ldots, \{X_N\}\}$, $\mathbf{S}^{(K)} = \{X\}$ a každý rozklad $\mathbf{S}^{(t)}$ je zjemněním rozkladu $\mathbf{S}^{(t+1)}$.

Základním předpokladem úspěšného použití shlukovacích metod je, že objekty mají tendenci seskupovat se do shluků a nemají charakter více méně homogenního chaosu. Ověření tohoto předpokladu bývá obtížné a není mu vždy věnována náležitá pozornost. Dalším důležitým problémem je volba vhodného shlukovacího postupu. Obecné pravidlo, jak zvolit pro daný problém optimální postup, neexistuje a volba bývá často subjektivní. Významnou roli zde hraje studium vlastností jednotlivých postupů a využití některých kriterií pro hodnocení získaných shluků z hlediska jejich kompaktnosti a vzájemné isolovanosti.

Problematice shlukové analysy byl věnován seminář, který probíhal v oddělení teorie pravděpodobnosti a matematické statistiky Matematického ústavu ČSAV. Na seminář navázala Letní škola shlukové analysy, kterou uspořádala v roce 1976 matematická vědecká sekce Jednoty československých matematiků a fyziků a liberecká pobočka JČSMF. Letní škola ukázala, že metody shlukové analysy se v Československu na různé úrovni na řadě pracovišť používají a vedou v mnoha případech k velmi dobrým výsledkům. Současně se projevila velká nejednotnost v používané terminologii. To vyplývá ze skutečnosti, že o shlukové analyse nebyly dosud uveřejněny v české matematické literatuře žádné práce. Samotný termín cluster se vyskytuje v češtině v různých podobách jako hnízdo, shluk, svazek, trs apod. Název *shluk* se nám jevil jako nejvhodnější mimo jiné proto, že od slova shluk lze snadno tvořit slova odvozená; můžeme např. hovořit o *shlukovacích* postupech a rozklad $\boldsymbol{S}$ či pokrytí $\boldsymbol{M}$ množiny objektů nazvat *shluknutím* a jejich vytváření nazývat *shlukováním.*

Úkolem tohoto článku je přehledně zachytit základní myšlenky shlukové analysy a přispět tak k dalšímu rozšíření jejích metod, informovat o literatuře a pokusit se o sjednocení české terminologie. V první části pojednáme o mírách nepodobnosti (respektive podobnosti) objektů a shluků a funkcionálech kvality rozkladu. Druhá část se zabývá jednotlivými typy shlukovacích postupů. Část třetí je věnována vlastnostem těchto postupů.

1. MÍRY NEPODOBNOSTI A FUNKCIONÁLY KVALITY ROZKLADU

Nechť je dána množina N objektů $X = \{X_1, X_2, \ldots, X_N\}$. Úkolem shlukové analysy je nalézt rozklad $\boldsymbol{S} = \{S_1, S_2, \ldots, S_n\}$ množiny X do n shluků tak, aby objekty patřící do téhož shluku si byly v jistém smyslu podobné a objekty patřící do různých shluků se svými vlastnostmi co nejvíce lišily. V některých případech požadavek vzájemné disjunktnosti shluků neodpovídá skutečné situaci. Proto úlohu někdy zobecňujeme v tom smyslu, že shluky S_i se mohou vzájemně překrývat, a místo rozkladu hledáme pokrytí množiny X shluky S_i, tj. takovou množinu $\boldsymbol{S} = \{S_1, S_2, \ldots$ $\ldots, S_n\}$, aby $X = \bigcup_{i=1}^{n} S_i$.

Podobnost či odlišnost objektů X_i posuzujeme na základě p určitých znaků, jejichž hodnoty na objektech změříme a jež jsou zvoleny podle konkrétní úlohy. Proto v dalším budeme ztotožňovat objekty X_i s vektory $(x_{i1}, \ldots, x_{ip})$, kde x_{ij} je hodnota naměřená na j-tém znaku i-tého objektu. (Pokud by na různých objektech byly naměřeny stejné vektory, mohli bychom je rozlišit třeba tak, že bychom formálně připojili další souřadnici označující index objektu.) Zpravidla lze vektory X_i považovat za prvky nějakého prostoru $\boldsymbol{X}$, který nazýváme základním prostorem. Často bývá $\boldsymbol{X}$ eukleidovský prostor $\mathscr{E}_p$ nebo alespoň lineární prostor.

Shlukovací metody jsou obvykle založeny na mírách nepodobnosti (resp. podobnosti) objektů a shluků, nebo na funkcionálech kvality rozkladu do shluků.

1.1. Míry nepodobnosti objektů. *Míra nepodobnosti* na X je nezáporná reálná funkce d definovaná na $X \times X$ taková, že $d(X_i, X_j) = d(X_j, X_i)$ pro všechna $X_i, X_j \in X$ a $d(X_i, X_i) = 0$, $i = 1, 2, \ldots, N$.

Míru nepodobnosti na X můžeme např. zavést tak, že nejprve definujeme míru nepodobnosti d na $\boldsymbol{X}$ a míru nepodobnosti na X definujeme jako restrikci míry nepodobnosti d na X. Tuto restrikci budeme nazývat indukovanou mírou nepodobnosti. Zřejmě každá metrika na $\boldsymbol{X}$ indukuje míru nepodobnosti na X.

Nechť $X = \{X_1, X_2, \ldots, X_N\}$ a d je míra nepodobnosti na X. Z vlastností míry nepodobnosti vyplývá, že d je jednoznačně určena svými $\binom{N}{2}$ hodnotami $d(X_i, X_j)$ pro $i < j$. Můžeme proto d považovat za prvek $\binom{N}{2}$-rozměrného eukleidovského prostoru $\mathscr{E}_{\binom{N}{2}}$. Tato representace míry nepodobnosti d je užitečná, jestliže chceme definovat vzdálenost dvou měr nepodobnosti.

Míru nepodobnosti d na X nazýváme *ultrametrickou*, jestliže d splňuje tzv. *ultrametrickou nerovnost*

$$d(X_i, X_j) \leqq \max\left[d(X_i, X_k), d(X_j, X_k)\right] \quad \text{pro každé} \quad X_i, X_j \quad \text{a} \quad X_k \in X\,.$$

V tabulce 1 uvádíme některé míry nepodobnosti definované na $\mathscr{E}_p$, které jsou často užívány k zavedení měr nepodobnosti na X v případě, že hodnoty znaků měřených na objektech jsou vesměs reálná čísla.

O matici $\boldsymbol{W}$ v míře nepodobnosti d_M předpokládáme, že je positivně definitní. $\boldsymbol{W}$ bývá volena jako

$$\boldsymbol{W} = \sum_{i=1}^{N} (X_i - \bar{X})(X_i - \bar{X})'\,,$$

kde $\bar{X} = (1/N)\sum_{i=1}^{N} X_i$, za předpokladu $|\boldsymbol{W}| \neq 0$. V takovém případě se $\boldsymbol{W}$ někdy nazývá matice rozsevu.

Míry nepodobnosti d_1 a d_2 jsou zvláštním případem míry d_m. Další míry nepodobnosti jsou uvedeny v [5] a [3].

Tabulka 1: Některé míry nepodobnosti

Název	Tvar
l_1-metrika	$d_1(X_i, X_j) = \sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert$
Eukleidovská metrika	$d_2(X_i, X_j) = [\sum_{k=1}^{p} (x_{ik} - x_{jk})^2]^{1/2}$
l_m-metrika	$d_m(X_i, X_j) = [\sum_{k=1}^{p} \lvert x_{ik} - x_{jk} \rvert^m]^{1/m}$
Sup-metrika	$d_\infty(X_i, X_j) = \max_k \{\lvert x_{ik} - x_{jk} \rvert\}$
Mahalanobisova zobecněná metrika	$d_M(X_i, X_j) = (X_i - X_j)' \, \mathbf{W}^{-1} (X_i - X_j)$
Úhel sevřený X_i a X_j	$d_s(X_i, X_j) = \arccos \dfrac{\sum_{k=1}^{p} x_{ik} x_{jk}}{[\sum_{k=1}^{p} x_{ik}^2 \sum_{k=1}^{p} x_{jk}^2]^{1/2}}$
Úhel sevřený $X_i - \bar{X}_i$ a $X_j - \bar{X}_j$	$d_r(X_i, X_j) = \arccos r_{ij}$, kde $r_{ij} = \dfrac{\sum_{k=1}^{p} (x_{ik} - \bar{X}_i)(x_{jk} - \bar{X}_j)}{[\sum_{k=1}^{p} (x_{ik} - \bar{X}_i)^2 \sum_{k=1}^{p} (x_{jk} - \bar{X}_j)^2]^{1/2}}$, $\bar{X}_i = \dfrac{1}{p} \sum_{k=1}^{p} x_{ik}$

Míra nepodobnosti je někdy zaváděna pomocí tzv. potenciálové funkce $K(X_i, X_j)$ definované na $X \times X$, a to vzorcem

$$d_K(X_i, X_j) = \sqrt{(K(X_i, X_i) + K(X_j, X_j) - 2K(X_i, X_j))} .$$

Je-li $X \subset \mathcal{E}_p$ a volíme-li $K(X_i, X_j) = (X_i, X_j) = \sum_{k=1}^{p} x_{ik} x_{jk}$, dostaneme míru nepodobnosti d_2. Potenciálovou funkci můžeme definovat prostřednictvím některé přirozené míry nepodobnosti d na X jako nerostoucí funkci d, např.

$$K(X_i, X_j) = e^{-\alpha d^2(X_i, X_j)}, \quad \alpha > 0 ,$$

nebo

$$K(X_i, X_j) = [1 + \alpha d^2(X_i, X_j)]^{-1}, \quad \alpha > 0.$$

Míry nepodobnosti na X je někdy vhodné kombinovat, což můžeme provést následujícím způsobem: jsou-li $d_1, d_2, \ldots, d_k$ míry nepodobnosti na X a je-li $\boldsymbol{V}$ symetrická positivně semidefinitní matice typu $k \times k$, potom $d = \boldsymbol{d}' \boldsymbol{V} \boldsymbol{d}$ je zřejmě míra nepodobnosti na X, kde $\boldsymbol{d}$ je vektor $\boldsymbol{d} = (d_1, d_2, \ldots, d_k)'$. Matici $\boldsymbol{V} = (v_{ij})$ můžeme interpretovat jako matici vah v_{ii} a podobností v_{ij} $(i \neq j)$ měr $d_1, d_2, \ldots, d_k$.

Duálním pojmem k míře nepodobnosti je *míra podobnosti*. Míra podobnosti na X je reálná funkce s na $X \times X$ taková, že $0 \leqq s(X_i, X_j) = s(X_j, X_i) \leqq 1$ pro všechna $X_i, X_j \in X$ a $s(X_i, X_i) = 1$. Každé míře podobnosti s na X je přiřazena míra nepodobnosti d_s na X vztahem $d_s(X_i, X_j) = 1 - s(X_i, X_j)$.

Míry podobnosti jsou používány ponejvíce v případě binárních znaků. Označme

n_{IJ} ... počet znaků k, pro které $x_{ik} = x_{jk} = 1$,

n_{ij} ... počet znaků k, pro které $x_{ik} = x_{jk} = 0$,

n_{iJ} ... počet znaků k, pro které $x_{ik} = 0,\ x_{jk} = 1$,

n_{Ij} ... počet znaků k, pro které $x_{ik} = 1,\ x_{jk} = 0$,

Často užívanými mírami podobnosti jsou následující míry:

$$n_{IJ}/(n_{IJ} + n_{Ij} + n_{ij}),$$

$$(n_{IJ} + n_{ij})/p,$$

$$n_{IJ}/p,$$

$$2n_{IJ}/(2n_{IJ} + n_{Ij} + n_{iJ}),$$

$$2(n_{IJ} + n_{ij})/(p + n_{IJ} + n_{ij}),$$

$$n_{IJ}/[n_{IJ} + 2(n_{Ij} + n_{iJ})],$$

$$(n_{IJ} + n_{ij})/(p + n_{Ij} + n_{iJ}).$$

Je-li s míra podobnosti a f neklesající funkce taková, že $f(0) \geqq 0, f(1) = 1$, potom $f(s)$ je zřejmě mírou podobnosti na X. Jestliže f je nerostoucí funkce a $f(1) = 0$, pak $f(s)$ je míra nepodobnosti na X.

1.2. Míry nepodobnosti shluků. Při mnohých shlukovacích metodách, převážně pak u tzv. hierarchických metod, o nichž bude pojednáno v části 2, shlukujeme objekty z X v 1. kroku shlukovacího algoritmu na základě míry nepodobnosti objektů; vytvořené shluky považujeme za nové „objekty", které v dalších krocích opět shlukujeme na základě měr nepodobnosti shluků. Míry nepodobnosti shluků jsou obvykle definovány pomocí měr nepodobnosti objektů. Některé míry nepodobnosti shluků jsou uvedeny v tabulce 2, kde d značí míru nepodobnosti na X, a S_k, S_m jsou prvky rozkladu $\mathbf{S}$. Symbol $|S|$ značí v celém článku počet prvků množiny S.

Tabulka 2: Míry nepodobnosti shluků

Název	Tvar
Vzdálenost nejbližších prvků shluků S_k a S_m (nejbližší soused)	$\tilde{d}_1(S_k, S_m) = \min_{X_i \in S_k, X_j \in S_m} \{d(X_i, X_j)\}$
Vzdálenost nejvzdálenějších prvků shluků S_k a S_m (nejvzdálenější soused)	$\tilde{d}_2(S_k, S_m) = \max_{X_i \in S_k, X_j \in S_m} \{d(X_i, X_j)\}$
Průměrná vzdálenost prvků shluků S_k a S_m	$\tilde{d}_3(S_k, S_m) = \frac{1}{\lvert S_k \rvert \cdot \lvert S_m \rvert} \sum_{X_i \in S_k} \sum_{X_j \in S_m} d(X_i, X_j)$
Vzdálenost průměrů shluků (centroidní)	$\tilde{d}_c(S_k, S_m) = d(\bar{S}_k, \bar{S}_m)$, kde $\bar{S}_k = \frac{1}{\lvert S_k \rvert} \sum_{X_i \in S_k} X_i$
Kolmogorovova zobecněná vzdálenost	$\tilde{d}^r(S_k, S_m) = \left[\frac{1}{\lvert S_k \rvert \lvert S_m \rvert} \sum_{X_i \in S_k} \sum_{X_j \in S_m} (d(X_i, X_j))^r \right]^{1/r}$

Míra nepodobnosti shluků $\tilde{d}^r$ zahrnuje jako zvláštní případ míry nepodobnosti $\tilde{d}_1$ resp. $\tilde{d}_2$ resp. $\tilde{d}_3$ při $r \to -\infty$ resp. $r \to +\infty$ resp. $r = 1$.

1.3. Funkcionály kvality rozkladu. Funkcionál kvality rozkladu je reálná funkce $f(\mathbf{S})$ definovaná na množině všech rozkladů $\mathbf{S}$ množiny objektů X. Úlohu shlukování pak můžeme formulovat jako úlohu nalézt extrém vhodně zvoleného funkcionálu $f(\mathbf{S})$ na množině všech rozkladů nebo na některé její podmnožině. Funkcionál kvality rozkladu by měl vystihovat naše apriorní představy o optimálním shluknutí v konkrétní situaci.

Uveďme některé běžně užívané funkcionály kvality rozkladu (d je míra nepodobnosti na X):

$$Q_1(\mathbf{S}) = \sum_{m=1}^{n} \sum_{X_i \in S_m} d^2(X_i, \bar{S}_m), \quad \text{kde} \quad \bar{S}_m = (1/\lvert S_m \rvert) \sum_{X_i \in S_m} X_i,$$

$$Q_2(\mathbf{S}) = \sum_{m=1}^{n} \sum_{X_i, X_j \in S_m} d^2(X_i, X_j),$$

$Q_3(\mathbf{S}) = \det\left(\sum_{m=1}^{n} |S_m| \mathbf{W}_m\right)$, kde $\mathbf{W}_m$ je výběrová kovariační matice shluku S_m,

$Q_4(\mathbf{S}) = \prod_{m=1}^{n} (\det \mathbf{W}_m)^{|S_m|}$,

$Q_5(\mathbf{S}) = I_1(\mathbf{S}) + I_2(\mathbf{S})$, kde $I_1(\mathbf{S}) = \sum_{m=1}^{n} \sum_{X_i \in S_m} d(X_i, \bar{S}_m)$ nebo $I_1(\mathbf{S}) = Q_1(\mathbf{S})$

a $I_2(\mathbf{S}) = c \,.\, n = c \,.\, |\mathbf{S}|$, kde $c > 0$ je konstanta.

Funkcionály $Q_1(\mathbf{S}) - Q_4(\mathbf{S})$ se používají při rozkladech na pevný počet shluků $n < N$ nebo pro n splňující $n_1 \leqq n \leqq n_2 < N$. Funkcionál $Q_5(\mathbf{S})$ je používán, není-li počet shluků n předem dán, a $I_2(\mathbf{S})$ hraje roli ztrátové funkce závisející na počtu shluků.

A. N. KOLMOGOROV předložil obecné schéma, používající dvou funkcionálů, z nichž první je mírou koncentrace objektů v rozkladu $\mathbf{S}$ a druhý střední mírou podobnosti objektů ve shlucích.

Míra koncentrace je dána vztahem

$$Z_s(\mathbf{S}) = \left[\frac{1}{N} \sum_{i=1}^{N} \left(\frac{v(X_i)}{N}\right)^s\right]^{1/s},$$

kde $v(X_i)$ je počet objektů ve shluku obsahujícím objekt X_i. Volba čísla s závisí na typu řešené úlohy shlukování. Pro některé hodnoty s má funkcionál $Z_s(\mathbf{S})$ tvar:

$$\log Z_0(\mathbf{S}) = \sum_{m=1}^{n} \frac{|S_m|}{N} \log \frac{|S_m|}{N},$$

$$Z_{-1}(\mathbf{S}) = \frac{1}{n},$$

$$Z_1(\mathbf{S}) = \frac{1}{N} \sum_{j=1}^{n} \frac{v(X_j)}{N} = \frac{1}{N^2} \sum_{m=1}^{n} |S_m|^2,$$

$$Z_{-\infty}(\mathbf{S}) = \min_{1 \leqq m \leqq n} \left(\frac{|S_m|}{N}\right),$$

$$Z_{\infty}(\mathbf{S}) = \max_{1 \leqq m \leqq n} \left(\frac{|S_m|}{N}\right).$$

Poznamenejme, že pro libovolné s má míra koncentrace minimální hodnotu $1/N$, dosaženou pro rozklad X na jednoprvkové shluky a maximální hodnotu 1, kterou nabývá při spojení všech objektů do jednoho shluku.

Střední míra podobnosti objektů uvnitř shluků je dána vztahem

$$I^r(\mathbf{S}) = \left[\frac{1}{N} \sum_{m=1}^{n} |S_m| \left(\bar{d}^r(S_m, S_m)\right)^r\right]^{1/r},$$

kde $\tilde{d}^r$ je Kolmogorovova zobecněná vzdálenost, čili

$$I^r(\mathbf{S}) = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{1}{|S(X_i)|}\sum_{X_j \in S(X_i)}(d(X_i, X_j))^r\right]^{1/r},$$

kde $S(X_i)$ je shluk obsahující objekt X_i.

Funkcionály kvality rozkladu na shluky jsou pak voleny jako funkce $f(Z_s(\mathbf{S}), I^r(\mathbf{S}))$.

Zcela obecně uveďme, že většinu funkcionálů $f(\mathbf{S})$ lze vyjádřit ve tvaru

$$f(\mathbf{S}) = f(u_1, \ldots, u_k),$$

kde $u_i = u_i(\mathbf{S})$ jsou funkcionály vyjadřující různé vlastnosti, které očekáváme od optimálního rozkladu na shluky, a charakterizující strukturu uvnitř shluků nebo mezi shluky, např.:

u_1 ... stupeň podobnosti objektů uvnitř shluků,
u_2 ... stupeň odlišnosti shluků,
u_3 ... homogenita rozložení objektů uvnitř shluků,
u_4 ... rovnoměrnost rozložení objektů do n shluků apod.

2. TYPY SHLUKOVACÍCH METOD

Mezi dosud navrženými metodami shlukování můžeme vysledovat tři základní typy metod konstrukce shluků, a to metody hierarchické, metody paralelní a metody sekvenční. Většinu známých shlukovacích metod můžeme začlenit pod některý z těchto typů.

Při *hierarchických metodách* shlukování se vytváří posloupnost $\mathbf{S}^{(t)}$ $(t = 1, 2, \ldots, K)$ rozkladů množiny objektů X taková, že rozklad $\mathbf{S}^{(t)}$ je zjemněním rozkladu $\mathbf{S}^{(t')}$ pro $t < t'$, $(t, t' = 1, 2, \ldots, K)$. Obvykle je rozklad $\mathbf{S}^{(K)}$ tvořen jediným shlukem, množinou X, a rozklad $\mathbf{S}^{(1)}$ systémem jednoprvkových množin, z nichž každá obsahuje právě jeden objekt z X. Příslušnost objektů shlukům v posloupnosti rozkladů lze vyjádřit tzv. dendrogramem. Hierarchie vytváření navzájem nepodobných skupin podobných si objektů je základním požadavkem v biologické taxonomii. Proto hierarchické shlukovací metody jsou pomocným nástrojem poznání v této disciplině a hlavním předmětem studia tzv. numerické taxonomie.

Paralelní shlukovací metody jsou nehierarchické iterační metody, které při každém iteračním kroku využívají pro konstrukci shluků všech shlukovaných objektů $X_i \in X$. Jejich cílem obvykle je určit globální nebo alespoň lokální extrém nějakého funkcionálu kvality rozkladu. Mezi paralelní metody lze např. zařadit metodu postupného přenosu objektu ze shluku do shluku a postupy založené na tzv. vzorových množinách.

Sekvenční metody jsou iterační shlukovací metody, jimiž jsou při každém iteračním kroku vytvářeny shluky na vybrané vlastní podmnožině množiny shlukovaných objektů. Výběr podmnožiny může být realizován např. náhodným výběrem prvků z X. V každém iteračním kroku je množina objektů vybraných ke shlukování rozšířením množiny objektů shlukovaných v kroku předcházejícím.

Jelikož při paralelních shlukovacích metodách je v každém iteračním kroku zpracovávána veškerá informace obsažená v měřeních na shlukovaných objektech, jsou paralelní metody při strojovém zpracování velmi náročné na kapacitu paměti a strojový čas počítačů. Tím je omezena praktická použitelnost paralelních metod, které jsou většinou aplikovány na shlukování nevelkého (vzhledem k parametrům počítače) počtu objektů. Naproti tomu jsou sekvenční metody svými vlastnostmi určeny pro shlukování velkého počtu objektů.

2.1. Hierarchické shlukovací metody. Předpokládejme, že pro objekty z X je dána míra jejich nepodobnosti d. Označme $\mathcal{C}(X)$ množinu všech měr nepodobnosti na X.

Posloupnost rozkladů $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(K)}$ množiny X nazveme hierarchickou, je-li rozklad $\mathcal{S}^{(t)}$ zjemněním rozkladu $\mathcal{S}^{(t')}$ pro $t < t', (t, t' = 1, 2, \ldots, K)$.

Hierarchickou metodu můžeme chápat jako metodu přiřazení prvků množiny $\mathcal{P}(X)$ všech konečných hierarchických posloupností rozkladů množiny X měrám nepodobnosti z $\mathcal{C}(X)$. V tomto smyslu je hierarchická metoda zobrazení $\mathcal{C}(X) \to \mathcal{P}(X)$.

Rozklad množiny objektů X na disjunktní shluky je jednoznačně určen ekvivalencí na X, tj. reflexivní, symetrickou a transitivní relací r definovanou na X. Relace r je podmnožina $X \times X$. Relace je

reflexivní, jestliže $(X_i, X_i) \in r$ pro všechna $X_i \in X$,
symetrická, jestliže z $(X_i, X_j) \in r$ plyne $(X_j, X_i) \in r$,
transitivní, jestliže z $(X_i, X_j) \in r$ a $(X_j, X_k) \in r$ plyne $(X_i, X_k) \in r$.

Je-li r ekvivalence, pak množiny $S_{X_i} = \{X_j : (X_i, X_j) \in r\}$, $i = 1, 2, \ldots, N$, tvoří rozklad množiny X. Naopak, je-li $\mathcal{S} = \{S_1, S_2, \ldots, S_n\}$ rozklad množiny X, je $r = \bigcup_{i=1}^{n} S_i \times S_i$ ekvivalencí na X.

Nechť $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}, \ldots, \mathcal{S}^{(K)}$ je hierarchická posloupnost rozkladů. Zapišme $\mathcal{S}^{(t)}$, $t = 1, 2, \ldots, K$, ve tvaru $\mathcal{S}^{(t)} = \{S_1^{(t)}, S_2^{(t)}, \ldots, S_{n_t}^{(t)}\}$ a označme $r(\mathcal{S}^{(t)}) = \bigcup_{i=1}^{n_t} S_i^{(t)} \times S_i^{(t)}$ ekvivalenci určenou rozkladem $\mathcal{S}^{(t)}$. Zřejmě je $r(\mathcal{S}^{(t)}) \subset r(\mathcal{S}^{(t')})$, pro $t < t'$. Předpokládáme-li, že $\mathcal{S}^{(K)}$ je tvořen jediným shlukem, jímž je množina všech objektů (tj. $r(\mathcal{S}^{(K)}) = X \times X$), je možno vyjádřit hierarchickou posloupnost rozkladů tzv. dendrogramem rozkladu X.

Označme $\mathcal{E}(X)$ množinu všech ekvivalencí na X. *Dendrogram rozkladu* X je funkce $c : [0, \infty) \to \mathcal{E}(X)$, která má vlastnosti:

a) je-li $0 \leqq h < h'$, pak $c(h) \subset c(h')$,
b) existuje h tak, že $c(h) = X \times X$,
c) ke každému h existuje $\delta > 0$ tak, že $c(h + \delta) = c(h)$.

Nechť c je dendrogram. Definujme na $X \times X$ funkci U_c vztahem

$$U_c(X_i, X_j) = \inf \{h : (X_i, X_j) \in c(h)\} . \tag{1}$$

Potom ekvivalence $c(h)$ je pro $h \in [0, \infty)$ určena vztahem

$$c(h) = \{(X_i, X_j) : U_c(X_i, X_j) \leqq h\} . \tag{2}$$

Snadno lze dokázat, že funkce U_c je ultrametrická míra nepodobnosti na X. Množinu všech ultrametrických měr nepodobnosti na X označíme $\boldsymbol{U}(X)$. Platí $\boldsymbol{U}(X) \subset \boldsymbol{C}(X)$.

Z (1) a (2) plyne, že zobrazení $c \to U_c$ je jednojednoznačné. Konečné hierarchické posloupnosti rozkladů z $\mathscr{S}(X)$ jsou tudíž určeny ultrametrickými měrami nepodobnosti na X. *Hierarchickou shlukovací metodu* budeme proto definovat jako zobrazení $D : \boldsymbol{C}(X) \to \boldsymbol{U}(X)$.

Nejužívanějšími hierarchickými metodami jsou *metoda nejbližšího souseda* (D_1), *metoda nejvzdálenějšího souseda* (D_2), *metoda nevážených průměrů* (D_3) a *vážených průměrů* (D_4). Při těchto metodách je počáteční rozklad $\boldsymbol{S}^{(1)}$ tvořen jednoprvkovými množinami $\{X_i\}$, $X_i \in X$, $i = 1, 2, \ldots, N$. Do téhož shluku rozkladu $\boldsymbol{S}^{(2)} = \{S_1^{(2)}, S_2^{(2)}, \ldots, S_{n_2}^{(2)}\}$ jsou zařazeny všechny objekty, jejichž vzájemná nepodobnost je rovna $m = \min_{X_i, X_j \in X} d(X_i, X_j)$. Tedy pro $X_i \neq X_j$ a $X_i, X_j \in S_k^{(2)}$, $k = 1, 2, \ldots, n_2$, platí $d(X_i, X_j) = m$. Pro takto vytvořené shluky je způsobem popsaným dále stanovena míra nepodobnosti shluků a postup se opakuje tak, že shluky považujeme za objekty při dalším kroku shlukování. Proces se ukončí, když všechny objekty X vytvoří jediný shluk.

Předpokládejme, že v t-tém kroku vytvoří shluky $S_{k_1}^{(t)}, S_{k_1}^{(t)}, \ldots, S_{k_r}^{(t)}$ shluk $S_k^{(t+1)}$ a shluky $S_{m_1}^{(t)}, S_{m_2}^{(t)}, \ldots, S_{m_s}^{(t)}$ shluk $S_m^{(t+1)}$. Při metodách D_1, D_2, D_3, D_4 jsou míry nepodobnosti shluků určovány takto:

$$D_1 \colon \tilde{d}_1(S_k^{(t+1)}, S_m^{(t+1)}) = \min_{(i,j)} \tilde{d}_1(S_{k_i}^{(t)}, S_{m_j}^{(t)}) ,$$

$$D_2 \colon \tilde{d}_2(S_k^{(t+1)}, S_m^{(t+1)}) = \max_{(i,j)} \tilde{d}_2(S_{k_i}^{(t)}, S_{m_j}^{(t)}) ,$$

$$D_3 \colon \tilde{d}_3(S_k^{(t+1)}, S_m^{(t+1)}) = r^{-1}s^{-1} \sum_i \sum_j \tilde{d}_3(S_{k_i}^{(t)}, S_{m_j}^{(t)}) ,$$

$$D_4 \colon \tilde{d}_4(S_k^{(t+1)}, S_m^{(t+1)}) = |S_k^{(t+1)}|^{-1} |S_m^{(t+1)}|^{-1} \sum_i \sum_j |S_{k_i}^{(t)}| \, |S_{m_j}^{(t)}| \, \tilde{d}_4(S_{k_i}^{(t)}, S_{m_j}^{(t)}) .$$

Míry nepodobnosti shluků pro metody D_1, D_2, D_3 jsou po řadě shodné s měrami $\tilde{d}_i$ tabulky 2, $(i = 1, 2, 3)$.

Nejstarší z těchto metod, metodu nejbližšího souseda, která je někdy nazývána dendritovou metodou [7], můžeme charakterizovat následujícím způsobem. Je-li $d, d' \in \boldsymbol{C}(X)$, nazveme míru nepodobnosti d *dominantní* vzhledem k d' a označíme $d' \leqq d$, jestliže $d'(X_i, X_j) \leqq d(X_i, X_j)$ pro všechna $X_i, X_j \in X$. Metoda nejbližšího

souseda je zobrazení $D : \mathbf{C}(X) \to \mathbf{U}(X)$, které měrám nepodobnosti $d \in \mathbf{C}(X)$ přiřazuje prvky $D(d) \in \mathbf{U}(X)$ takové, že $D(d)$ je největší prvek z $\mathbf{U}(X)$, pro který $D(d) \leqq d$.

Další hierarchickou metodou je *centroidní metoda* (D_c), při které předpokládáme, že objekty $X_i \in X$ lze representovat jako prvky $X_i \in \boldsymbol{X}$, kde základní prostor $\boldsymbol{X}$ je lineární. Nechť je na $\boldsymbol{X}$ dána míra nepodobnosti d, která indukuje míru nepodobnosti na X. Při centroidní metodě je míra nepodobnosti shluků volena jako míra $\tilde{d}_c$ tabulky 2.

Definici hierarchické shlukovací metody lze zobecnit tak, aby zahrnula též případ nedisjunktního pokrytí množiny objektů X jejími podmnožinami. Množinu $\boldsymbol{M} = \{M_1, M_2, \ldots, M_n\}$ podmnožin $M_i \subset X$ nazveme pokrytím X, je-li $\bigcup_{i=1}^{n} M_i = X$. Konečnou posloupnost pokrytí $\boldsymbol{M}^{(1)}, \boldsymbol{M}^{(2)}, \ldots, \boldsymbol{M}^{(K)}$ $(\boldsymbol{M}^{(t)} = \{M_1^{(t)}, M_2^{(t)}, \ldots, M_{n_t}^{(t)}\}$, $t = 1, 2, \ldots, K$, $M_j^{(t)} \subset X)$ nazveme hierarchickou, je-li pokrytí $M^{(t)}$ zjemněním pokrytí $\boldsymbol{M}^{(t')}$ pro $t < t'$, $t, t' = 1, 2, \ldots, K$.

Označíme-li $\mathscr{M}(X)$ množinu hierarchických posloupností pokrytí (jejíž prvky splňují přirozené požadavky specifikované níže), potom hierarchickou metodu můžeme v širším smyslu chápat jako zobrazení $\mathbf{C}(X) \to \mathscr{M}(X)$.

Od metody shlukování požadujeme, aby vytvářela pouze taková pokrytí, která mají rozumné vlastnosti. Z úvah budeme zřejmě chtít vyloučit např. taková zobrazení $\mathbf{C}(X) \to \mathscr{M}(X)$, která některé míře nepodobnosti $d \in \mathbf{C}(X)$ přiřazují hierarchickou posloupnost pokrytí, v níž existuje prvek $\boldsymbol{M}^{(t)} = \{M_1^{(t)}, M_2^{(t)}, \ldots, M_{n_t}^{(t)}\}$ a $i \neq j$ tak, že $M_i^{(t)} \subset M_j^{(t)}$. Množinu $\mathscr{M}(X)$ proto vymezíme podmínkou, aby byla tvořena jen takovými hierarchickými posloupnostmi pokrytí, jejichž prvky mají v některém smyslu dobrou strukturu překrývání.

Dobrou strukturu překrývání můžeme zavést např. pomocí tzv. maximálně vázaných množin. Nechť r je symetrická a reflexivní relace na $\boldsymbol{X}$. Množinu $M \subset X$ nazveme *maximálně vázanou* vzhledem k relaci r, je-li $M \times M \subseteq r$ (a tedy $r \cap M \times M$ je ekvivalence na M) a jestliže pro každé $X_i \in X - M$ existuje $X_j \in M$ tak, že $(X_i, X_j) \notin r$. Každé pokrytí $\boldsymbol{M} = \{M_1, M_2, \ldots, M_n\}$ množiny X určuje relaci $r(\boldsymbol{M}) = \bigcup_{i=1}^{n} M_i \times M_i$. Pokrytí $\boldsymbol{M}$ nazveme *pokrytím s dobrou strukturou překrývání*, s tliže M_i jsou maximálně vázané množiny vzhledem k relaci $r(\boldsymbol{M})$.

Množinu $\mathscr{M}(X)$ pak definujeme jako množinu všech hierarchických posloupností pokrytí s dobrou strukturou překrývání. V dalším předpokládáme, že $\mathscr{M}(X)$ je takto definována.

Je-li r symetrická, reflexivní relace na X, pak množina $\boldsymbol{M}(r) = \{M_1(r), \ldots, M_{n_r}(r)\}$ všech maximálně vázaných množin vzhledem k r tvoří pokrytí X s dobrou strukturou překrývání. Mezi reflexivními, symetrickými relacemi na X a pokrytími s dobrou strukturou překrývání je tudíž přirozený jednojednoznačný vztah. Přitom pro $r_1 \subset r_2$ platí, že pokrytí $\boldsymbol{M}(r_1)$ je zjemněním pokrytí $\boldsymbol{M}(r_2)$.

Hierarchické posloupnosti pokrytí z $\mathscr{M}(X)$ můžeme vyjádřit pomocí tzv. *funkce stratifikovaného shlukování.*

Označme $\Sigma(X)$ množinu všech reflexivních, symetrických relací na X. Funkci $c : [0, \infty) \to \Sigma(X)$, která splňuje podmínky a), b), c) požadované pro dendrogram nazýváme funkcí stratifikovaného shlukování. Funkce stratifikovaného shlukování je zobecněním dendrogramu; u dendrogramu jsou $c(h)$ ekvivalence na X, kdežto u funkce stratifikovaného shlukování jsou $c(h)$ reflexivní, symetrické relace na X.

Nechť c je funkce stratifikovaného shlukování, potom definujeme-li na $X \times X$ funkci U_c vztahem (1), je U_c míra nepodobnosti na X (tj. $U_c \in \mathbf{C}(X)$). Naopak, je-li U_c míra nepodobnosti na X, potom je vztahem

$$c(h) = \{(X_i, X_j) : U_c(X_i, X_j) \leqq h\}\,, \quad h \in [0, \infty)\,, \tag{3}$$

určena funkce stratifikovaného shlukování a tedy hierarchická posloupnost pokrytí $z\ \mathscr{M}(X)$.

Mezi funkcemi stratifikovaného shlukování a měrami nepodobnosti tudíž existuje přirozený jednojednoznačný vztah. Hierarchickou metodu v širším smyslu můžeme proto definovat jako zobrazení $\mathbf{C}(X) \to \mathbf{C}(X)$. Hierarchické metody v širším smyslu přiřazují měrám nepodobnosti $d \in \mathbf{C}(X)$ hierarchické posloupnosti pokrytí s dobrou strukturou překrývání. V tomto smyslu jsou hierarchické. Protože však v literatuře se hierarchickými nazývají pouze takové metody shlukování, které lze vyjádřit dendrogramem a graficky znázornit hierarchickým stromem, a tuto vlastnost hierarchické metody v širším smyslu obecně nemají, budeme je nazývat stratifikovanými metodami shlukování.

Nechť $\mathbf{A} \subseteq \mathbf{C}(X)$, $\mathbf{Z} \subseteq \mathbf{C}(X)$. *Stratifikovanou metodu shlukování* definujeme jako zobrazení $D : \mathbf{A} \to \mathbf{Z}$.

Triviální metodou shlukování při $\mathbf{A} = \mathbf{Z}$ je metoda $D(d) = d$, pro $d \in \mathbf{A}$. Protože metoda shlukování objektů by měla koncentrovat informaci, která je obsažena v míře nepodobnosti $d \in \mathbf{A}$, je vhodné požadovat $\mathbf{A} \supset \mathbf{Z}$. Množina měr nepodobnosti $\mathbf{A}$ je někdy nazývána *množinou dat* a množina $\mathbf{Z}$ *cílovou množinou.*

Cílová množina $\mathbf{Z}$ může být volena tak, aby stratifikovaná metoda shlukování vytvářela pokrytí množiny X s dobrou strukturou překrývání a s některými dalšími vlastnostmi struktury překrývání. Volíme-li např. $\mathbf{Z} = \mathbf{U}(X)$, potom stratifikovaná metoda shlukování $D : \mathbf{A} \to \mathbf{U}(X)$ je hierarchickou metodou shlukování, tj. metodou vytvářející disjunktní shluky. Vzhledem k jednojednoznačnému vztahu mezi relacemi $r \in \Sigma(X)$ a pokrytími X s dobrou strukturou překrývání můžeme další vlastnosti struktury překrývání formulovat jako vlastnosti relací. Volba struktury překrývání je tak volbou podmnožiny $\Sigma^*(X) \subseteq \Sigma(X)$. Stratifikované metody shlukování, které vytvářejí pokrytí se strukturou překrývání $\Sigma^*(X)$, jsou právě takové metody, které přiřazují měrám nepodobnosti funkce stratifikovaného shlukování $c^* : [0, \infty) \to \Sigma^*(X)$. Označme $\mathbf{Z}^* = \mathbf{Z}^*(X)$ tu část $\mathbf{C}(X)$, která odpovídá takovým funkcím c^*. Potom každá stratifikovaná metoda shlukování, která vytváří pouze pokrytí se strukturou překrývání $\Sigma^*(X)$, je zobrazení $D : \mathbf{A} \to \mathbf{Z}^*$.

Požadujeme-li kupříkladu, aby stratifikovaná metoda shlukování vytvářela jen taková pokrytí s dobrou strukturou překrývání, jejichž množiny se překrývají nejvýše v $(k-1)$ objektech, lze dokázat, že všechna taková pokrytí jsou právě určena množinou $\Sigma_k^*(X)$ všech tzv. slabě k-transitivních relací.

Relaci $r \in \Sigma(X)$ nazýváme *slabě k-transitivní*, jestliže z $\{\{X_i\} \times S \cup S \times S \cup S \times \times \{X_j\}\} \subseteq r$ plyne $(X_i, X_j) \in r$ pro všechna $X_i, X_j \in X$ a všechna $S \subset X$ taková, že $|S| = k$.

Cílová množina $\boldsymbol{Z}_k^* \subset \boldsymbol{C}(X)$, která odpovídá podmínce struktury překrývání $\Sigma_k^*(X)$, je množina všech slabých k-ultrametrických měr nepodobnosti na X.

Míru nepodobnosti $d \in \boldsymbol{C}(X)$ nazýváme *slabě k-ultrametrickou*, jestliže pro každé $S \subset X$, $|S| = k$ a libovolná $X_i, X_j \in X$ platí tzv. *slabě k-ultrametrická nerovnost*:

$$d(X_i, X_j) \leqq \max \{d(Y, Z) : Y \in S \cup \{X_i, X_j\},\ Z \in S\} .$$

Všechny stratifikované shlukovací metody, které vytvářejí pokrytí se strukturou překrývání $\Sigma_k^*(X)$, jsou právě všechna zobrazení $D : \boldsymbol{A} \to \boldsymbol{Z}_k^*$. Pro $k = 1$ je $\boldsymbol{Z}_k^* = \boldsymbol{U}(X)$.

V porovnání se slabou k-transitivitou je silnější podmínkou kladenou na strukturu překrývání silná k-transitivita relací ze $\Sigma(X)$.

Relaci $r \in \Sigma(X)$ nazýváme *silně k-transitivní*, jestliže z $\{\{X_i\} \times S \cup S \times \{X_j\}\} \subseteq \subseteq r$ plyne $(X_i, X_j) \in r$ pro všechna $S \subset X$ a $|S| = k$, $X_i, X_j \in X$.

Označíme-li $\Sigma_k^{**}(X)$ množinu všech silně k-transitivních relací na X, je $\Sigma_k^{**}(X) \subset \subset \Sigma_k^*(X)$. Všechny shlukovací metody, které vytvářejí pokrytí se strukturou překrývání $\Sigma_k^{**}(X)$, jsou všechna zobrazení typu $D : \boldsymbol{A} \to \boldsymbol{Z}_k^{**}$, kde $\boldsymbol{Z}_k^{**}$ je množinou všech silně k-ultrametrických měr nepodobnosti na X.

Míru nepodobnosti $d \in \boldsymbol{C}(X)$ nazýváme *silně k-ultrametrickou*, jestliže pro každou množinu $S \subset X$, $|S| = k$ a všechna $X_i, X_j \in X$ platí tzv. *silně k-ultrametrická nerovnost*

$$d(X_i, X_j) \leqq \max \{d(Y, Z) : Y \in \{X_i, X_j\},\ Z \in S\} .$$

Zřejmě $\boldsymbol{Z}_k^{**} \subset \boldsymbol{Z}_k^*$.

Stratifikované metody shlukování jsou oproti metodám hierarchickým lepším modelem shlukování v těch reálných situacích, kdy měření x_{ij} na objektech X_i jsou prováděna s relativně velkou chybou, která má náhodnou povahu. Tehdy je možno X_i (při $X_i \in \mathscr{E}_p$), $i = 1, 2, \ldots, N$, považovat za realizace náhodných veličin. Vzhledem k nepřesnosti měření je pak přirozenější požadovat, aby metoda shlukování vytvářela pokrytí a nikoli jen rozklady množiny objektů X. Stratifikované metody shlukování vytvářející pokrytí se strukturou překrývání $\Sigma^*(X)$ resp. $\Sigma^{**}(X)$ jsou prvním krokem aproximace takové reálné situace. Otevřenou zůstává otázka konstrukce algoritmu těchto metod. Doposud navržené algoritmy stratifikovaných metod shlukování se strukturou překrývání $\Sigma^*(X)$ resp. $\Sigma^{**}(X)$ jsou velmi komplikované a neefektivní pro výpočet i v případě jednoduchých stratifikovaných metod [9].

2.2. Paralelní a sekvenční shlukovací metody. Nejčastěji užívanými paralelními metodami jsou metoda vzorových množin a metoda přenosu objektu ze shluku do shluku.

Nechť $X = \{X_1, X_2, \ldots, X_N\}$ je množina shlukovaných objektů. Budiž $m_1, m_2, \ldots$ $\ldots, m_n$ n-tice přirozených čísel, $\sum_{i=1}^{n} m_i < N$. Systém $\boldsymbol{E} = \{E_1, E_2, \ldots, E_n\}$ disjunktních podmnožin $E_i \subset X$, které mají postupně mohutnosti $m_1, m_2, \ldots, m_n$, nazýváme *systémem vzorových množin.* V případě $m_1 = m_2 = \ldots = m_n = 1$ hovoříme o *systému vzorových bodů.*

Algoritmy *shlukování metodou vzorových* množin jsou založeny na funkcích φ, ψ. $\varphi(X_i, A)$ je funkce definovaná pro $X_i \in X$, $A \subset X$. Interpretujeme ji jako míru možnosti vyjádření objektu X_i množinou A. $\psi(X_i, A)$ je rovněž definována pro $X_i \in X$, $A \subset X$. Interpretujeme ji však jako míru možnosti vyjádření množiny A objektem X_i. Je-li d míra nepodobnosti na X, můžeme funkce φ, ψ volit např. takto:

$$\varphi(X_i, A) = \sum_{Y \in A} d(X_i, Y), \quad \psi(X_i, A) = \varphi(X_i, A).$$

Při shlukování metodou vzorových množin vyjdeme z jistým způsobem zvoleného počátečního systému vzorových množin, kterému pomocí funkce φ přiřadíme systém shluků (rozklad množiny X). Takto vzniklému systému shluků přiřadíme pomocí funkce ψ nový systém vzorových množin a postup opakujeme. Proces ukončíme, když v sousledných iteračních krocích dostaneme stejný rozklad.

Předpokládejme, že počet shluků, které má metoda vytvořit, je $\leqq n$. Nechť $\boldsymbol{E}^{(0)} = \{E_1^{(0)}, E_2^{(0)}, \ldots, E_n^{(0)}\}$ je počáteční systém vzorových množin. Systému $\boldsymbol{E}^{(0)}$ přiřadíme rozklad $\boldsymbol{S}^{(0)} = \boldsymbol{S}(\boldsymbol{E}^{(0)}) = \{S_1(\boldsymbol{E}^{(0)}), S_2(\boldsymbol{E}^{(0)}), \ldots, S_n(\boldsymbol{E}^{(0)})\}$ množiny objektů X, pomocí funkce φ, vztahem

$$S_i^{(0)} = S_i(\boldsymbol{E}^{(0)}) = \{Y \in X : \varphi(Y, E_i^{(0)}) < \varphi(Y, E_j^{(0)}); \quad i \neq j, j = 1, 2, \ldots, n\}, \quad i = 1, 2, \ldots, n. \tag{4}$$

Do $S_i^{(0)}$ jsou tedy zařazovány ty objekty, které jsou nejlépe representovány množinou $E_j^{(0)}$. Jestliže pro některá i, j, $i \neq j$ a $Y \in X$ je $\min_k \varphi(Y, E_k^{(0)}) = \varphi(Y, E_i^{(0)}) = \varphi(Y, E_j^{(0)})$, zařazujeme prvek Y do shluku s nejmenší hodnotou indexu.

K $\boldsymbol{S}^{(0)}$ přiřadíme pomocí funkce ψ systém vzorových množin $\boldsymbol{E}^{(1)} = \boldsymbol{E}(\boldsymbol{S}^{(0)}) = \{E_1(\boldsymbol{S}^{(0)}), E_2(\boldsymbol{S}^{(0)}), \ldots, E_{n_1}(\boldsymbol{S}^{(0)})\}$. Nechť $F_i(\boldsymbol{S}^{(0)}) = \{Y \in X : \psi(Y, S_i^{(0)}) < \psi(Y, S_j^{(0)}); i \neq j, j = 1, 2, \ldots, n\}$, $i = 1, 2, \ldots, n$. V případě rovnosti zařazujeme Y do množiny s nejmenší hodnotou indexu. Je-li $|F_i(\boldsymbol{S}^{(0)})| \leqq m_i$, položíme $E_i^{(1)} = E_i(\boldsymbol{S}^{(0)}) = F_i(\boldsymbol{S}^{(0)})$. Je-li $|F_i(\boldsymbol{S}^{(0)})| > m_i$, zařadíme do $E_i(\boldsymbol{S}^{(0)})$ m_i objektů $Y \in F_i(\boldsymbol{S}^{(0)})$ s nejmenšími hodnotami funkce $\psi(Y, S_i^{(0)})$.

Postup opakujeme a ukončíme v t-tém kroku, jestliže $\boldsymbol{E}^{(t)} = \boldsymbol{E}^{(t-1)}$, tj. $\boldsymbol{S}^{(t)} = \boldsymbol{S}^{(t-1)}$.

Konvergence metody a vlastnosti rozkladu, který metodou získáme, zřejmě závisí na volbě funkcí φ, ψ. Nechť $\boldsymbol{E}, \tilde{\boldsymbol{E}}$ jsou dva systémy vzorových množin a $\boldsymbol{S}(\tilde{\boldsymbol{E}})$ rozklad přiřazený pravidlem (4) systému vzorových množin $\tilde{\boldsymbol{E}}$. Označme $\psi(E_{\pi_i}, S_i(\tilde{\boldsymbol{E}}))$ součet $\sum_{Y \in E_{\pi_i}} \psi(Y, S_i(\tilde{\boldsymbol{E}}))$, kde $\pi = (\pi_i)$ je permutace množiny indexů $1, 2, \ldots, n$.

$\psi(E_{\pi_i}, S_i(\tilde{\boldsymbol{E}}))$ je mírou možnosti vyjádření vzorové množiny E_{π_i} množinou $S_i(\tilde{\boldsymbol{E}})$.

Potom

$$\Delta(\boldsymbol{E}, \tilde{\boldsymbol{E}}) = \min_{\pi} \sum_{i} \psi(E_{\pi_i}, S_i(\tilde{\boldsymbol{E}}))$$

je globální mírou možnosti vyjádření systému vzorových množin $\boldsymbol{E}$ rozkladem $\boldsymbol{S}(\tilde{\boldsymbol{E}})$.

Vyjadřuje-li rozklad $\boldsymbol{S}(\tilde{\boldsymbol{E}})$ nějaký systém vzorových množin $\boldsymbol{E}$ lépe než systém $\tilde{\boldsymbol{E}}$, z kterého vznikl, je přirozené požadovat, aby $\boldsymbol{E}$ bylo vyjádřeno shluky $\boldsymbol{S}(\boldsymbol{E})$ lépe než shluky $\boldsymbol{S}(\tilde{\boldsymbol{E}})$.

Proto požadujeme, aby funkce φ, ψ splňovaly podmínku

(5) $$\text{z} \quad \Delta(\tilde{\boldsymbol{E}}, \tilde{\boldsymbol{E}}) > \Delta(\boldsymbol{E}, \tilde{\boldsymbol{E}}) \quad \text{plyne} \quad \Delta(\boldsymbol{E}, \tilde{\boldsymbol{E}}) > \Delta(\boldsymbol{E}, \boldsymbol{E}) .$$

Systém vzorových množin $\boldsymbol{E} = \{E_1, E_2, \ldots, E_n\}$ je *lokálně optimální*, jestliže pro všechna $Y \in X - E_i$, $Z \in E_i$ platí

$$\psi(Y, S_i(\boldsymbol{E})) \geqq \psi(Z, S_i(\boldsymbol{E})) \quad \text{pro} \quad i = 1, 2, \ldots, n .$$

Jestliže funkce φ, ψ splňují podmínku (5) a funkce φ je taková, že

(6) $$\varphi(X_i, S) \neq \varphi(X_i, S') \quad \text{pro} \quad S, S' \subset X , \quad S \cap S' = \emptyset \quad \text{a všechna} \quad X_i \in X ,$$

potom $\boldsymbol{E}^{(t)} \to \bar{\boldsymbol{E}}$, kde $\bar{\boldsymbol{E}}$ je lokálně optimální systém vzorových množin [4]. Podmínka (5) je zřejmě splněna např. pro $\psi = \varphi$.

Budiž $\boldsymbol{E}$ systém vzorových množin. Položme $\psi = \varphi$, kde φ splňuje (6), a definujme

$$Q(S_i) = \sum_{Y \in E_i} \varphi(Y, S_i) , \quad i = 1, 2, \ldots, n ,$$

kde $S_i = \{Y \in X : \varphi(Y, E_i) < \varphi(Y, E_j); j \neq i, j = 1, 2, \ldots, n\}$.

Potom $H(\boldsymbol{S}(\boldsymbol{E})) = \sum_i Q(S_i)$ můžeme považovat za hodnotu funkcionálu kvality rozkladu $\boldsymbol{S}(\boldsymbol{E})$; v dalším značíme $H(\boldsymbol{E}) = H(\boldsymbol{S}(\boldsymbol{E}))$. V tomto případě $\boldsymbol{E}^{(t)} \to \bar{\boldsymbol{E}}$ a $H(\bar{\boldsymbol{E}})$ je lokální minimum funkcionálu $H(\boldsymbol{E})$. Ve speciálním případě, kdy $\boldsymbol{E} = \{e_1, e_2, \ldots, e_n\}$ je množina[1]) vzorových bodů, lze při volbě funkce $\varphi = d^2(X_j, e_i)$, $X_j \in X$, $e_i \in \boldsymbol{E}$, kde d je míra nepodobnosti na X, funkcionál $H(\boldsymbol{E})$ psát ve tvaru $H_1(\boldsymbol{E}) = \sum_{i=1}^{n} \sum_{X_j \in S_i} d^2(X_j, e_i)$. Je-li míra nepodobnosti d na X taková, že $d(X_i, X_j) = 0$ právě

[1]) Pro zjednodušení zápisu ztotožníme zde i v dalším jednobodové množiny $\{e_i\}$ s body $e_i \in X$.

tehdy, když $X_i = X_j$ a dále, že pro $(X_i, X_j) \neq (Y_i, Y_j)$, $X_i \neq X_j$ platí $d(X_i, X_j) \neq d(Y_i, Y_j)$, potom paralelní postup založený na systému vzorových bodů vede k lokální minimalizaci funkcionálu $H_1(\mathbf{E})$.

Poznamenejme ještě, že postup lze zobecnit na případ volby prvků vzorových množin ze základního prostoru $\boldsymbol{X}$. Je-li $\boldsymbol{X}$ lineární prostor, pak můžeme např. volit $\varphi(X_i, A) = d(X_i, \bar{X}(A))$, kde $\bar{X}(A) = (1/|A|) \sum_{X_i \in A} X_i$.

Někdy je též vhodné definovat funkci ψ, kterou jsou přiřazovány systémy vzorových množin $\mathbf{E}^{(t)}$ rozkladům $\mathbf{S}^{(t)}$, indukcí tak, aby závisela na systému vzorových množin $\mathbf{E}^{(t-1)}$, z něhož shluky $S_j^{(t)}$ vznikly, tj. $\psi = \psi(X_i, S_j^{(t)}(\mathbf{E}^{(t-1)}))$.

Příkladem je funkce

$$\psi(X_i, S_j(\mathbf{E})) = \frac{\varphi(X_i, S_j(\mathbf{E}))\, \varphi(X_i, E_j)}{(\sum_k \varphi(X_i, E_k))^2},$$

kde $\mathbf{E}$ je systém vzorových množin a $\mathbf{S}(\mathbf{E}) = \{S_1(\mathbf{E}), S_2(\mathbf{E}), \ldots, S_n(\mathbf{E})\}$ rozklad tomuto systému přiřazený pravidlem (4).

Paralelní postup shlukování se obvykle aplikuje na několik počátečních systémů vzorových množin a z výsledných rozkladů se vybírá ten, který má v jistém smyslu nejlepší strukturu. Je-li struktura např. charakterizována funkcionálem, vybereme ten rozklad, pro který je hodnota funkcionálu kvality rozkladu největší (resp. nejmenší).

Postup lze zobecnit na případ, kdy počet shluků ve vytvářených rozkladech není omezen číslem $n < N$. V tomto případě se postupuje tak, že v prvním kroku algoritmu:

1) Určí se n_0, $0 < n_0 < N$, přirozená čísla $m_1, m_2, \ldots, m_{n_0}$ a počáteční systém vzorových množin $\mathbf{E}^{(0)} = \{E_1^{(0)}, E_2^{(0)}, \ldots, E_{n_0}^{(0)}\}$, kde $|E_i^{(0)}| = m_i$.

2) Dále se zvolí tzv. prahová hodnota φ_0 funkce φ. Pomocí funkce $\varphi(X_i, A)$ se systému vzorových množin $\mathbf{E}^{(0)}$ přiřadí rozklad, jehož prvky tvoří množiny

$$(7) \qquad S_i^{(0)} = \{Y \in X : \varphi(Y, E_i^{(0)}) < \varphi(Y, E_j^{(0)});\ j \neq i,\ j = 1, 2, \ldots, n_0\} \cap \cap \{Y \in X : \varphi(Y, E_i^{(0)}) \leqq \varphi_0\},\quad i = 1, 2, \ldots, n_0,$$

a dále jednoprvkové množiny $\{X_k\}$, kde X_k jsou objekty, které nejsou vztahem (7) zařazeny do množin $S_i^{(0)}$.

3) Vytvořené shluky postupně spojujeme pomocí funkce $\tilde{\psi}(S_i, S_j)$, která je mírou podobnosti shluků S_i, S_j a může být volena např. takto:

$$\tilde{\psi}(S_i, S_j) = \frac{1}{2}\left[\frac{1}{|S_i|} \sum_{Y \in S_i} \psi(Y, S_j) + \frac{1}{|S_j|} \sum_{Y \in S_j} \psi(Y, S_i)\right].$$

Zvolíme prahovou hodnotu $\tilde{\psi}_0$ funkce $\tilde{\psi}(S_i, S_j)$ a spojíme shluky, pro které $\tilde{\psi}(S_i, S_j)$ je minimální a menší než $\tilde{\psi}_0$. Postup spojování shluků opakujeme a ukončíme, když pro všechny dvojice shluků v rozkladu je funkce $\tilde{\psi} \geqq \tilde{\psi}_0$.

Postup spojování zřejmě nemusí být jednoznačný. Naznačme, jak shlukům vytvořeným spojením dvou shluků přiřazujeme vzorové množiny: jednobodové shluky

$\{X_k\}$ považujeme za shluky vytvořené jednobodovou vzorovou množinou X_k a spojení $S_i \cup S_j$ shluků S_i, S_j, které byly vytvořeny vzorovými množinami E_i, E_j, přiřadíme vzorovou množinou E_{ij}, do níž zařadíme $m_{ij} = \max(|E_i|, |E_j|)$ prvků $Y \in X$, na kterých nabývá funkce $\psi(Y, S_i \cup S_j)$ nejmenších hodnot. Tak vznikne systém vzorových množin $\boldsymbol{E}^{(1)}$ pro další iterační krok algoritmu.

Metody shlukování založené na vzorových bodech může být v některých případech užito k lokální optimalizaci funkcionálů kvality rozkladu (např. funkcionálů typu $H(\boldsymbol{E})$).

Metodou, která je užívána výhradně k optimalizaci funkcionálů kvality rozkladu, je *metoda přenosu* objektu ze shluku do shluku. Při této metodě předpokládáme, že objekty z X jsou uspořádány v posloupnost $X_1, X_2, \ldots, X_N$. Metoda je založena na iteračním algoritmu. V prvním kroku vycházíme z počátečního rozkladu $\boldsymbol{S}^{(0)} = \{S_1^{(0)}, S_2^{(0)}, \ldots, S_n^{(0)}\}$ a bod X_1 se postupně přenáší do všech shluků. Vzniknou tak rozklady $\boldsymbol{S}_{1i}^{(0)}$, $i = 1, 2, \ldots, n$. Z nich vybereme rozklad $\boldsymbol{S}_1^{(0)}$, pro který je hodnota $Q(\boldsymbol{S}_{1i}^{(0)})$ daného funkcionálu $Q(\boldsymbol{S})$ minimální. V rozkladu $\boldsymbol{S}_1^{(0)}$ přenášíme bod X_2 ze shluku do shluku a ze vzniklých rozkladů vybereme rozklad $\boldsymbol{S}_2^{(0)}$ s minimální hodnotou funkcionálu Q, atd. První krok algoritmu uzavře přenášení bodu X_N, které vede k výběru rozkladu $\boldsymbol{S}^{(1)} = \boldsymbol{S}_N^{(0)}$, který je výchozím rozkladem dalšího kroku algoritmu.

Metoda zřejmě závisí na počátečním rozkladu $\boldsymbol{S}^{(0)}$ a na uspořádání množiny objektů X v posloupnost.

S paralelní metodou shlukování založenou na vzorových množinách je příbuzná sekvenční *metoda průměrů.*

Předpokládejme $X_i \in \mathscr{E}_p$. Metoda průměrů vytváří postupně systém n vzorových bodů $\boldsymbol{E}^{(t)} = \{e_1^{(t)}, e_2^{(t)}, \ldots, e_n^{(t)}\}$, kterému je funkcí $\varphi(X_i, A)$ přiřazován rozklad $\boldsymbol{S}^{(t)} = \boldsymbol{S}(\boldsymbol{E}^{(t)})$ pravidlem (4). Poznamenejme, že pro aplikaci pravidla (4) v případě vesměs jednobodových vzorových množin stačí funkci φ definovat jen pro jednobodové množiny $A \subset X$. Nechť je míra nepodobnosti d' na X indukována mírou nepodobnosti d na $\mathscr{E}_p$. Položme $\varphi(X_i, Y) = d(X_i, Y)$ pro $X_i \in X$, $Y \in \mathscr{E}_p$.

Metoda průměrů vychází z počátečního systému vzorových množin $\boldsymbol{E}^{(0)} = \{e_1^{(0)}, e_2^{(0)}, \ldots, e_n^{(0)}\}$, kterým jsou přiřazeny váhy $v_1^{(0)} = v_2^{(0)} = \ldots = v_n^{(0)} = 1$. Nechť $\boldsymbol{E}^{(t-1)} = \{e_1^{(t-1)}, e_2^{(t-1)}, \ldots, e_n^{(t-1)}\}$ je systém vzorových množin vytvořený v $(t-1)$ kroku a označme $v_1^{(t-1)}, v_2^{(t-1)}, \ldots, v_n^{(t-1)}$ postupně váhy vzorových bodů $e_1^{(t-1)}, e_2^{(t-1)}, \ldots, e_n^{(t-1)}$. V t-tém kroku vybereme (např. náhodně) z množiny X objekt X_t. Nechť $e_i^{(t-1)}$ je vzorový bod z $\boldsymbol{E}^{(t-1)}$, pro který $d(X_t, e_i^{(t-1)}) = \min_j d(X_t, e_j^{(t-1)})$. Potom $\boldsymbol{E}^{(t)} = \{e_j^{(t)}\}_{j=1}^n$ s váhami $\{v_j^{(t)}\}_{j=1}^n$, kde

$$e_i^{(t)} = \frac{e_i^{(t-1)} + X_t}{v_i^{(t-1)} + 1}, \quad v_i^{(t)} = v_i^{(t-1)} + 1$$

a

$$e_j^{(t)} = e_j^{(t-1)}, \quad v_j^{(t)} = v_j^{(t-1)} \quad \text{pro} \quad j \neq i, \quad j = 1, 2, \ldots, n.$$

Počáteční systém vzorových množin $\mathbf{E}^{(0)}$ se někdy vytváří náhodným výběrem bez vracení na X. Vlastnosti metody průměrů lze v případě $\varphi = d(X_i, e_i) = \varrho(X_i, e_i)$, kde ϱ je eukleidovská metrika na $\mathscr{E}_p$, charakterizovat pravděpodobnostně.

Předpokládejme, že $\mathscr{X}$ je p-rozměrná spojitá náhodná veličina a nechť jí indukovaná pravděpodobnostní míra P na $\mathscr{E}_p$ má vlastnost, že existuje uzavřená, omezená a konvexní množina $U^* \subset \mathscr{E}_p$, pro kterou $P(U^*) = \int_{U^*} \mathrm{d}P = 1$ a pro jejíž každou otevřenou podmnožinu $U \subset U^*$ s positivní lebesgueovskou mírou je $P(U) > 0$.

Nechť $S \subset \mathscr{E}_p$ a $P(S) > 0$. Označme $\mu(S) = (1/P(S)) \int_S Y \mathrm{d}P(Y)$ podmíněnou střední hodnotu veličiny $\mathscr{X}$ za podmínky S. Je-li $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$ rozklad prostoru $\mathscr{E}_p$ takový, že $P(S_i) > 0$, $i = 1, 2, \ldots, n$, označíme

$$\mu[\mathbf{S}] = (\mu(S_1), \mu(S_2), \ldots, \mu(S_n))$$

vektor podmíněných středních hodnot veličiny $\mathscr{X}$ vzhledem ke shlukům $S_1, S_2, \ldots, S_n$.

Budiž $\mathbf{E} = \{e_1, e_2, \ldots, e_n\}$ systém vzorových bodů a $\mathbf{S}(\mathbf{E}) = \{S_1(\mathbf{E}), S_2(\mathbf{E}), \ldots, S_n(\mathbf{E})\}$ rozklad prostoru $\mathscr{E}_p$ přiřazený systému $\mathbf{E}$ užitím pravidla (4) pro $Y \in \mathscr{E}_p$. Poznamenejme, že vzhledem ke spojitosti veličiny $\mathscr{X}$ má množina, na které je rozklad nejednoznačný, nulovou pravděpodobnost.

Množinu vzorových bodů nazýváme *nestrannou*, jestliže $\mu[\mathbf{S}(\mathbf{E})] = (e_1, e_2, \ldots, e_n)$. Položme

$$Q_1(\mathbf{E}) = \sum_{i=1}^{n} \int_{S_i(\mathbf{E})} \varrho^2(Y, \mu(S_i(\mathbf{E}))) \, P(\mathrm{d}Y)$$

a

$$\tilde{Q}_1(\mathbf{E}) = \sum_{i=1}^{n} \int_{S_i(\mathbf{E})} \varrho^2(Y, e_i) \, P(\mathrm{d}Y) \, .$$

Funkcionály $Q_1(\mathbf{E})$ a $\tilde{Q}_1(\mathbf{E})$ jsou funkcionály kvality rozkladu prostoru $\mathscr{E}_p$ a jsou zobecněním funkcionálu $Q_1(\mathbf{S})$. Funkcionál $\tilde{Q}_1(\mathbf{E})$ je zobecnění funkcionálu $H_1(\mathbf{E})$.

Nechť $Y_1, Y_2, \ldots, Y_n, \ldots$ je posloupnost nezávislých náhodných veličin, které mají rozložení pravděpodobnosti shodné s rozložením náhodné veličiny $\mathscr{X}$. Nechť při metodě průměrů je $\mathbf{E}^{(0)} = \{Y_1, Y_2, \ldots, Y_n\}$ a $X_t = Y_{n+t}$, $t = 1, 2, \ldots$. Veličina X_t má význam náhodného výběru objektu v t-tém kroku algoritmu.

Za těchto předpokladů s pravděpodobností 1 platí

$$\lim_{s \to \infty} \frac{1}{s+1} \sum_{t=0}^{s} \sum_{i=1}^{n} P(S_i^{(t)}) \, \varrho(\mu(S_i^{(t)}), e_i^{(t)}) = 0 \, .$$

Dále platí [12], že posloupnost náhodných veličin $\tilde{Q}_1(\mathbf{E}^{(t)})$ konverguje skoro všude a s pravděpodobností 1 je $\lim_{t \to \infty} \tilde{Q}_1(\mathbf{E}^{(t)}) = Q_1(\mathbf{E})$, kde $\mathbf{E}$ je nějakou nestrannou množinou vzorových bodů.

3. VLASTNOSTI SHLUKOVACÍCH POSTUPŮ

Z předchozí části vyplývá, že pro řešení dané úlohy shlukové analysy máme k disposici velké množství shlukovacích postupů. Stojíme tedy před problémem, který postup zvolit. Jak již bylo řečeno, ve shlukové analyse nejsou známa pravidla, která by umožňovala zvolit pro danou úlohu v jistém smyslu nejlepší shlukovací postup. K usnadnění volby shlukovacího postupu se definují různé žádoucí vlastnosti, které by rozumné shlukovací postupy měly mít, a pak se zkoumá, které shlukovací postupy tyto vlastnosti mají. Takové postupy se někdy nazývají přípustné shlukovací postupy vzhledem k dané vlastnosti ([6]). Budeme tohoto názvu používat, i když není zcela vhodný. Mohl by totiž svádět k domněnce, že nepřípustnost shlukovacího postupu značí jeho nevhodnost pro aplikace. Spíše však jde o to uvědomit si, jaké přednosti a nedostatky daný shlukovací postup má a zda jeho vlastnosti odpovídají požadavkům kladeným na řešení konkrétní úlohy shlukování. Přípustné shlukovací postupy můžeme definovat dvojím způsobem. V prvém případě klademe podmínky přímo na shlukovací postup. Často však nejprve definujeme přípustné rozklady a požadujeme, aby daný postup vytvářel pouze přípustné rozklady.

Je zcela přirozené požadovat, aby shlukovací postup použitý na množinu objektů, která má výraznou strukturu, tuto strukturu odhalil. Rozeznáváme tři druhy dobré struktury objektů. Objekty $X_1, X_2, \ldots, X_N$ mají *dokonalou strukturu*, jestliže existují shluky $S_1, S_2, \ldots, S_n$ a čísla $d_1 < d_2$ tak, že pro $X_i \in S_l$ a $X_j \in S_m$ platí $d(X_i, X_j) = d_1$ pro $l = m$ a $d(X_i, X_j) = d_2$ pro $l \neq m$, $l, m = 1, 2, \ldots, n$. Objekty $X_1, X_2, \ldots, X_N$ mají *dobrou n-strukturu*, jestliže existují shluky $S_1, S_2, \ldots, S_n$ tak, že $\max \{d(X_i, X_j) : X_i \in S_l,\ X_j \in S_l,\ l = 1, 2, \ldots, n\} < \min \{d(X_i, X_j) : X_i \in S_l,\ X_j \in S_m,\ j \neq m\}$. Konečně objekty $X_1, X_2, \ldots, X_N$ mají *přesnou strukturu dendrogramu*, jestliže míra nepodobnosti d splňuje ultrametrickou nerovnost, tj. $d \in \mathbf{U}(X)$. Shlukovací postup je *přípustný vzhledem k dokonalé struktuře*, respektive *vzhledem k dobré n-struktuře*, jestliže při použití na objekty s dokonalou, respektive dobrou n-strukturou vede k vytvoření příslušných shluků $S_1, S_2, \ldots, S_n$. Hierarchický shlukovací postup je *přípustný vzhledem k přesné struktuře dendrogramu*, jestliže při použití na objekty s přesnou strukturou dendrogramu vytvoří příslušný dendrogram. Pro stratifikované shlukovací postupy se zavádí vlastnost *přiměřenosti*. Postup $D : \mathbf{A} \to \mathbf{Z}$ je *přiměřeně přípustný*, jestliže $\emptyset \neq \mathbf{Z} \subset \mathbf{A}$ a $D : \mathbf{Z} = \mathrm{id}$, kde id je identické zobrazení $\mathbf{Z}$ na sebe. V případě, že $\mathbf{Z} = \mathbf{U}(X)$, je hierarchický shlukovací postup přiměřeně přípustný právě tehdy, když je přípustný vzhledem k přesné struktuře dendrogramu. Z hierarchických postupů jsou metoda nejbližšího souseda a metoda nejvzdálenějšího souseda přípustné jak vzhledem k dobré n-struktuře, tak vzhledem k přesné struktuře dendrogramu. Naproti tomu metoda centroidní je sice přípustná vzhledem k dobré n-struktuře, avšak není přípustná vzhledem k přesné struktuře dendrogramu. Ani metoda nejmenších čtverců ani metoda přenosu nejsou přípustné vzhledem k dobré n-struktuře. Zde i v dalším metodou nejmenších čtverců rozumíme nalezení globálního minima funkcionálu $Q_1(\mathbf{S}) = \sum_{l=1}^{n} \sum_{X_i \in S_l} d^2(X_i, \bar{S}_l)$. U

metody přenosu pak uvažujeme pouze speciální případ, kdy metodou přenosu hledáme lokální minimum téhož funkcionálu $Q_1(\boldsymbol{S})$.

Další skupina vlastností vychází z požadavku, aby při vynechání objektů či shluků nebo při jejich zdvojení vedl shlukovací postup ke stejnému výsledku. Nechť jsou dány objekty $X_1, X_2, \ldots, X_N$ a nechť shlukovací postup vytváří shluky $S_1, S_2, \ldots, S_n$. Některé objekty X_i vezmeme vícekrát a dostaneme tak objekty $Y_1, Y_2, \ldots, Y_M$; přitom počet opakování může být pro různé objekty různý. Shlukovací postup je *přípustný vzhledem ke zdvojování bodů*, jestliže při použití na objekty $Y_1, Y_2, \ldots, Y_M$ dostaneme stejné shluky jako pro objekty $X_1, X_2, \ldots, X_N$. Slabší podmínkou je *zdvojení shluků*, kdy postupujeme jako v předchozím případě, avšak počet opakování musí být pro všechny objekty v určitém shluku S_i stejný. Obě tyto vlastnosti jsou významné v případě že geometrický tvar shluků je důležitější než hustota objektů ve shlucích. Metoda nejbližšího souseda a metoda nejvzdálenějšího souseda jsou přípustné jak vzhledem ke zdvojování objektů tak vzhledem ke zdvojování shluků, kdežto metoda nejmenších čtverců a metoda přenosu nejsou přípustné vzhledem ani k jedné z těchto vlastností. Metoda centroidní je přípustná vzhledem ke zdvojování shluků, ale není přípustná vzhledem k zdvojování objektů.

Při *vynechání shluků* požadujeme, aby shlukovací postup, který pro objekty $X_1, X_2, \ldots, X_N$ vedl ke shlukům $S_1, S_2, \ldots, S_n$, vedl při použití na objekty z množiny $X - S_i$ k vytvoření shluků $S_1, S_2, \ldots, S_{i-1}, S_{i+1}, \ldots, S_n$. Metoda nejbližšího souseda, metoda nejvzdálenějšího souseda, centroidní metoda i metoda nejmenších čtverců jsou všechny přípustné vzhledem k vynechání shluků. Podobně lze definovat *přípustnou vzhledem k vynechání objektů*. Při tom požadujeme aby shlukovací postup který v množině objektů X vytvořil shluky $S_1\ S_2\ \ldots\ S_n$, vytvořil při použití na objekty z množiny $X - Y$ kde $Y \subset X$, shluky $S'_1, S'_2, \ldots, S'_m$ tak, že ke každému $j = 1, 2, \ldots, m$ existuje i_j takové, že $S'_j \subset S_{i_j}$, tj. vynechání některých objektů může nejvýše způsobit rozdělení původních shluků na shluky menší, nikoliv však jejich spojení či prolínání.

Velmi žádoucím se jeví požadavek, aby nebylo možno dosáhnout lepšího výsledku tím, že objekty jinak uspořádáme. Nechť $\boldsymbol{S} = \{S_1, S_2, \ldots, S_n\}$ je rozklad množiny objektů $X = \{X_1, X_2, \ldots, X_N\}$ a nechť $\pi : X \to X$ je permutace množiny objektů X. Definujme $\pi(\boldsymbol{S}) = \{\pi[S_1], \pi[S_2], \ldots, \pi[S_n]\}$. Rozklad $\boldsymbol{S}$ je *přípustný vzhledem k permutacím*, jestliže neexistuje permutace π tak, že rozklad $\pi(\boldsymbol{S})$ je stejnoměrně lepší v tomto smyslu: pro libovolné $X_i \in S_l$ a $X_j \in S_m$ platí $d(\pi(X_i), \pi(X_j)) \leqq d(X_i, X_j)$ pro $l = m$ a $d(\pi(X_i), \pi(X_j)) \geqq d(X_i, X_j)$ pro $l \neq m$ a přitom alespoň pro jednu dvojici indexů (i, j) platí ostrá nerovnost. Metoda nejbližšího souseda, metoda nejvzdálenějšího souseda a metoda nejmenších čtverců jsou všechny přípustné vzhledem k permutacím, kdežto metoda přenosu přípustná vzhledem k permutacím není. Obecně nelze očekávat, že lokální optimalizační postupy založené na počáteční volbě určitého rozkladu by byly přípustné vzhledem k permutacím.

V případě, že objekty jsou prvky lineárního prostoru, můžeme definovat konvexní přípustnost. Rozklad $\{S_1, S_2, \ldots, S_n\}$ je *konvexně přípustný*, jestliže $C(S_i) \cap C(S_j) =$

$= \emptyset$, kde $C(S_i)$ je konvexní obal shluku S_i. Mezi konvexně přípustné postupy patří metoda nejmenších čtverců a metoda přenosu, zatímco metoda nejbližšího souseda, metoda nejvzdálenějšího souseda a centroidní metoda nejsou konvexně přípustné. Konvexní přípustnost je motivována požadavkem, aby shluky byly jasně odděleny a nebyly do sebe zaklíněny. U této přípustnosti je dobře vidět, že není tak zcela jednoduché říci, které vlastnosti shlukových postupů jsou žádoucí. Požadavek, ze kterého definice vychází, je zcela přirozený. Na druhé straně mohou být objekty, které mají tendenci vytvářet shluky, jejichž konvexní obaly nejsou disjunktní. Pro takové objekty se konvexně přípustné shlukovací postupy naopak nehodí, protože shluky, které vytvářejí, jsou vždy konvexně přípustné.

Ve specielním případě, kdy objekty leží v rovině, můžeme podmínku konvexní připustnosti oslabit. Nechť je dán rozklad $\{S_1, S_2, \ldots, S_n\}$. Na množině S_i sestrojíme strom nejmenší možné délky (tj. s nejmenším součtem délek hran) a vzniklou souvislou množinu úseček označíme L_{S_i}. Postupujeme při tom tak, že objekty množiny S_i spojujeme metodou nejbližšího souseda. Rozklad $\{S_1, S_2, \ldots, S_n\}$ je *souvisle přípustný*, když $L_{S_i} \cap L_{S_j} = \emptyset$ pro $i \neq j$. Kromě metody nejmenších čtverců a metody přenosu je souvisle přípustná i metoda nejbližšího souseda. Metoda nejvzdálenějšího souseda a centroidní metoda však souvisle přípustné nejsou.

V případech, kdy se nemůžeme zcela spolehnout na velikost hodnot míry nepodobnosti a významné je spíše pořadí velikostí hodnot než hodnoty samy, je důležitý požadavek, aby výsledek shlukovacího postupu nezávisel na monotonní transformaci hodnot míry nepodobnosti. Nechť shlukovací postup použitý na objekty s mírou nepodobnosti d vede k rozkladu $\{S_1, S_2, \ldots, S_n\}$ a nechť $f : [0, \infty) \to [0, \infty)$ je rostoucí funkce taková, že $f(0) = 0$. Shlukovací postup se nazývá *monotónně přípustný*, jestliže jeho použití na stejné objekty s mírou nepodobnosti $f(d)$ vede k témuž rozkladu $\{S_1, S_2, \ldots, S_n\}$. Metoda nejbližšího souseda a metoda nejvzdálenějšího souseda jsou monotónně přípustné, kdežto centroidní metoda, metoda nejmenších čtverců a metoda přenosu nejsou monotonně přípustné. Slabší podmínku dostaneme, když uvažujeme pouze funkci $f(x) = ax$, kde $a > 0$. V tomto případě si zajistíme invariantnost vzhledem k lineární změně měřítka a shlukovací postup nazýváme *přípustným vzhledem k lineární změně měřítka.*

Konečně uvedeme některé definice přípustnosti pro stratifikované postupy. Nechť $D : \mathbf{A} \to \mathbf{Z}$ je stratifikovaný shlukovací postup a nechť platí: jestliže $d \in \mathbf{A}$, pak existuje $d' \in \mathbf{Z}$ tak, že $d' \leqq d$. Postup D je *přípustný vzhledem k zachovávání shluků*, jestliže platí: je-li M maximálně vázaná množina na úrovni h pro d (tj. maximálně vázaná vzhledem k relaci $\{(X_i, X_j) : d(X_i, X_j) \leqq h\}$), potom existuje maximálně vázaná množina M' na úrovni h pro $D(d)$ tak, že $M \subset M'$. Jinými slovy, maximálně vázané množiny na úrovni h na vstupu D zůstanou na úrovni h na výstupu D pohromadě. Formálně lze tento požadavek zapsat: $D(d) \leqq d$ pro všechny $d \in \mathbf{A}$.

Nechť $D : \mathbf{A} \to \mathbf{Z}$ a předpokládejme, že platí: je-li $\mathbf{Y}$ omezená podmnožina $\mathbf{Z}$, potom sup $\mathbf{Y} \in \mathbf{Z}$. Postup D je *optimálně přípustný*, jestliže platí: je-li $d' \in \mathbf{Z}$ a $D(d) \leqq \leqq d' \leqq d$, potom $D(d) = d'$. Vlastnost optimality zajišťuje, že ke koncentraci

informace dochází jen v nezbytně nutné míře a nejsou vytvářeny zbytečně velké shluky.

Zásadní důležitost má vlastnost spojitosti. Postup D je *spojitě přípustný*, jestliže zobrazení $D: \boldsymbol{A} \to \boldsymbol{Z}$ je spojité, chápeme-li množiny $\boldsymbol{A}$ a $\boldsymbol{Z}$ jako podmnožiny prostoru $\mathscr{E}_{\binom{N}{2}}$. Význam této podmínky spočívá v tom, že v jistém smyslu charakterisuje lokální stabilitu metody. Jde o to, aby malé změny v hodnotách míry nepodobnosti objektů vedly k malým změnám na výstupu shlukovacího postupu. Nemá-li shlukovací postup tuto vlastnost, pak je velmi citlivý na chyby měření znaků, které se odrážejí při stanovení hodnot míry nepodobnosti i na chyby způsobené zaokrouhlováním při výpočtech. Metoda nejbližšího souseda je spojitě přípustná avšak metoda nejvzdálenějšího souseda nikoliv.

Závěrem uveďme, že otevřenou otázkou zůstává hodnocení provedeného rozkladu. Skutečnost, že některý funkcionál kvality rozkladu nabývá pro získaný rozklad extrémní hodnoty zaručuje pouze, že rozklad množiny objektů X je vzhledem k námi volenému funkcionálu optimální, ale neříká nic o tom, zda se podařilo najít „skutečné" shluky, které v množině objektů X přirozeně existují. Intuitivně je zřejmé, že „skutečné" shluky by měly být pokud možno kompaktní a navzájem relativně izolované. Oba tyto požadavky zřejmě výrazně splňují objekty $X_1, X_2, \ldots, X_N$ s dobrou n-strukturou, pro které existuje rozklad $\{S_1, S_2, \ldots, S_n\}$ takový, že $\max \{d(X_i, X_j) : X_i \in S_l,\ X_j \in S_l,\ l = 1, 2, \ldots, n\} < \min \{d(X_i, X_j) : X_i \in S_l,\ X_j \in \in S_m,\ l \neq m\}$. Obecná kriteria, která by kompaktnost a relativní isolovanost shluků posoudila, zatím neexistují. Pro metodu nejbližšího souseda jsou známy některé výsledky na základě pravděpodobnostního přístupu ([11]). U hierarchických postupů metoda nejvzdálenějšího souseda minimalizuje maximální průměr shluků, kdežto metoda nejbližšího souseda maximalizuje minimální vzdálenost mezi shluky. Jestliže použití obou postupů vede k různým dendrogramům, pak zřejmě nelze splnit zároveň oba požadavky kompaktnosti a relativní isolovanosti shluků. Obecně se dá říci, že vedou-li různé hierarchické postupy k různým dendrogramům, je otázkou, zda se objekty hodí ke zpracování hierarchickým postupem.

Literatura

Problémům shlukování objektů bylo během posledních třiceti let věnováno mnoho prací publikovaných v řadě časopisů. V současnosti se projevuje snaha o syntézu výsledků a objevují se publikace knižní. Uveďme několik stručných poznámek ke čtyřem z nich.

Kniha [1] pojednává obecně o klasifikaci vícerozměrných pozorování. Třetí kapitola (59 stran) je věnována metodám shlukování objektů.

Publikace [2] podává dobrý přehled o problematice shlukové analýzy od klasifikace proměnných až po vyhodnocování shlukovacích metod. V rozsáhlé příloze (119 stran) autor uvádí řadu programů v jazyce FORTRAN IV.

Přehledná monografie [5] je psána heslovitým způsobem. Je věnována především hierarchickým metodám bez snahy o obecnější nadhled.

V knize [9] jsou zkoumány metody kondensace údajů, které se uplatňují především v problémech taxonomické klasifikace. Druhá část knihy (85 stran) je věnována hierarchickým a stratifikovaným metodám shlukové analysy; přitom důraz je kladen na přesné zavedení pojmů a matematickou teorii.

Všechny čtyři publikace obsahují obšírný seznam literatury; zejména v [5] nalezneme 409 položek.

[1] *С. А. Айвазян, З. И. Бежаева, О. В. Староверов:* Классификация многомерных наблюдений. Статистика, Москва, 1974. (240 stran).

[2] *M. R. Anderberg:* Cluster analysis for applications. Academic Press, New Nork, 1973. (359 stran).

[3] *R. M. Cormack:* A review of classification (with discussion). J. Royal Stat. Soc. Series A *134* (1971), 321—367.

[4] *E. Diday:* Une nouvelle méthode en classification automatique et reconnaissance des formes. La méthode des nuées dynamiques. Rev. Statist. Appl. *19* (1971), 19—34.

[5] *B. S. Duran, P. L. Odell:* Cluster analysis. A survey. Springer-Verlag, Berlin, 1974. (137 stran).

[6] *L. Fisher, J. W. van Ness:* Admissible clustering procedures. Biometrika *58* (1971), 91—104.

[7] *K. Florek, J. Łukaszewicz, H. Perkal, H. Steinhaus, S. Zubrzycki:* Sur la liaison et la division des points d'un ensemble fini. Coll. Math. *2* (1951), 282—285.

[8] *J. A. Hartigan:* Clustering algorithms. J. Wiley, New York, 1975.

[9] *N. Jardine, R. Sibson:* Mathematical taxonomy. J. Wiley, London, 1971. (286 stran).

[10] *M. G. Kendall:* Cluster analysis. In: Frontiers of pattern recognition. Academic Press, New York, 1972, 291—309.

[11] *R. F. Ling:* A probability theory of cluster analysis. J. Amer. Statist. Assoc. *68* (1973), 159—164.

[12] *J. MacQueen:* Some methods for classification and analysis of multivariate observation. Proc. Fifth Berkeley Sympos. Math. Statist. and Prob., (1) 1967, 281—297.

Adresa autorů: 115 67 Praha 1, Žitná 25 (Matematický ústav ČSAV).

Summary

CLUSTER ANALYSIS

ADOLF FILÁČEK, VÁCLAV KOUTNÍK, JIŘÍ VONDRÁČEK, Praha

This expository paper presents a survey of some methods of cluster analysis. The first part introduces dissimilarity measures for objects and clusters. In the second part hierarchical, parallel and sequential methods are discussed. The third part deals with the assessment of performance for clustering methods.