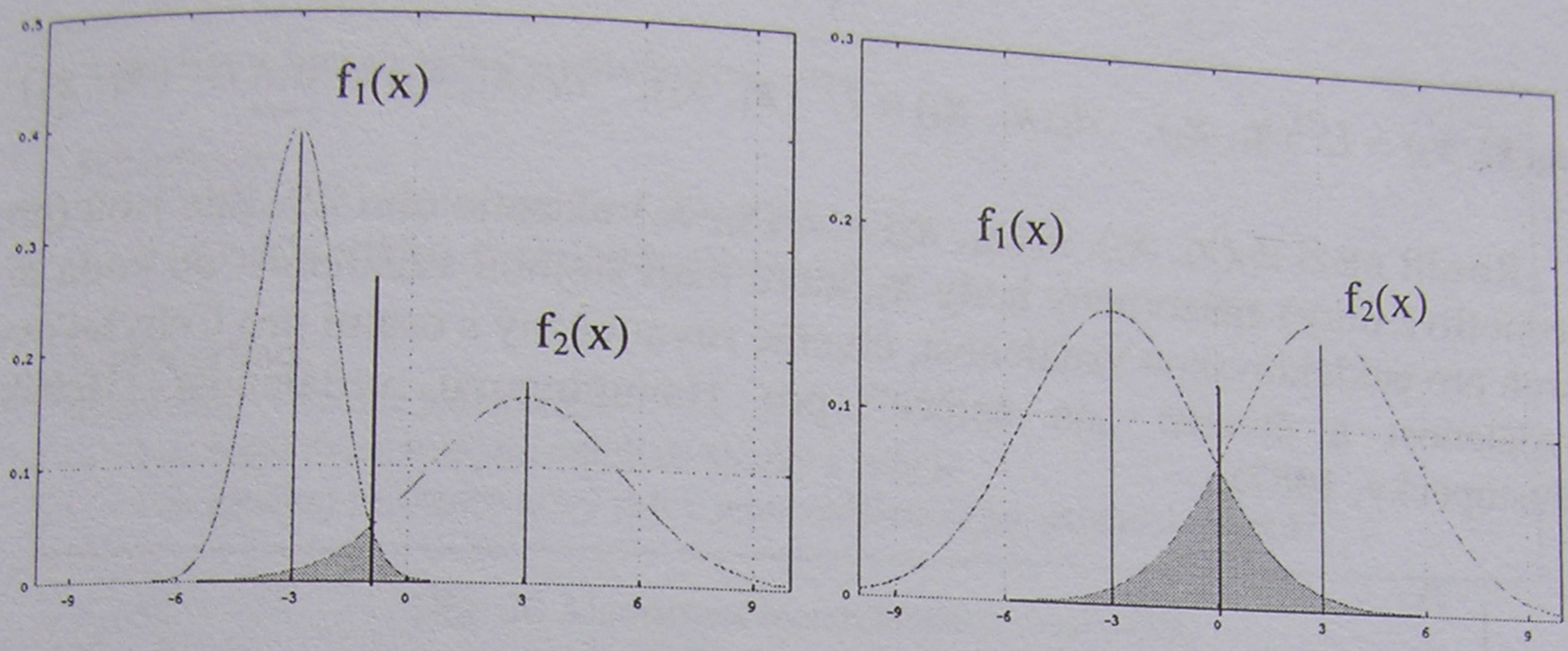
pravděpodobností $P(x|c_k)\ P(c_k)$ s různými rozptyly (vlevo) a stejnými rozptyly (vpravo). Pro stejné rozptyly tedy můžeme diskriminovat pouze na základě odhadů středních hodnot.



Obr. 28 Diskriminace podle minimální chyby, jednorozměrné rozdělení.

3.4 Shluková analýza

Shluková analýza hledá odpověď na otázku, zda lze pozorované příklady rozdělit do skupin (shluků) vzájemně si blízkých příkladů. Vychází se tedy z toho, že umíme měřit vzdálenost mezi příklady.

Předpokládejme, že každý příklad je charakterizován m numerickými veličinami. Vzdálenost mezi dvěma příklady $\boldsymbol{x}_1 = [x_{11},\ldots, x_{1m}]$ a $\boldsymbol{x}_2 =[x_{21},\ldots, x_{2m}]$ lze vyjádřit různými mírami. Uveďme zde např.:

- *Hammingovu vzdálenost*

$$d_{\mathrm{H}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sum_{j=1}^{m} |x_{1j} - x_{2j}|,$$

- *eukleidovskou vzdálenost*

$$d_{\mathrm{E}}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{j=1}^{m} (x_{1j} - x_{2j})^2},$$

- *Čebyševovu vzdálenost*

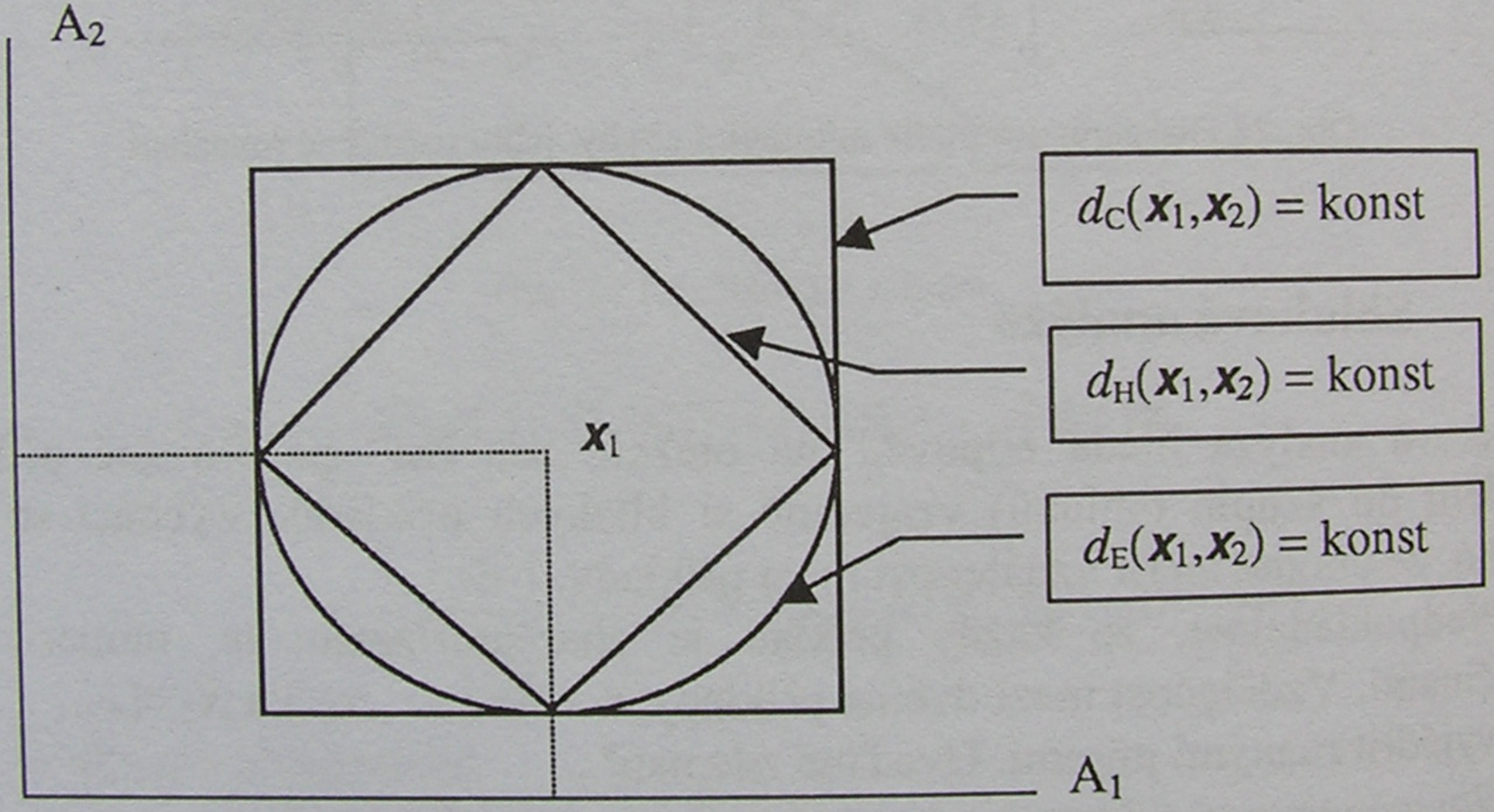
$$d_C = (\boldsymbol{x}_1, \boldsymbol{x}_2) = \max_i \left| x_{1j} - x_{2j} \right| .$$

Ve všech uvedených případech jde o speciální příklady Minkovského metriky

$$L^{(z)}(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt[z]{\sum_{j=1}^{m} (x_{1j} - x_{2j})^z} \ ;$$

$$d_H(\boldsymbol{x}_1, \boldsymbol{x}_2) = L^{(1)}(\boldsymbol{x}_1, \boldsymbol{x}_2), \quad d_E(\boldsymbol{x}_1, \boldsymbol{x}_2) = L^{(2)}(\boldsymbol{x}_1, \boldsymbol{x}_2), \quad d_C(\boldsymbol{x}_1, \boldsymbol{x}_2) = \lim_{z \to \infty} L^{(z)}(\boldsymbol{x}_1, \boldsymbol{x}_2) .$$

Rozdíl mezi $d_H(\boldsymbol{x}_1, \boldsymbol{x}_2)$, $d_E(\boldsymbol{x}_1, \boldsymbol{x}_2)$ a $d_C(\boldsymbol{x}_1, \boldsymbol{x}_2)$ ukazuje obr. 29. Zde jsou (pro jednotlivé míry) znázorněny body $\boldsymbol{x}_2$, které mají stejnou vzdálenost od bodu $\boldsymbol{x}_1$: kruh pro eukleidovskou vzdálenost, čtverec rovnoběžný s osami pro Čebyševovu vzdálenost a čtverec „na špičce“ pro Hammingovu vzdálenost (Hebák, Hustopecký, 1987).



Obr. 29 Body se stejnou vzdáleností od bodu *u*.

Uvedené míry vzdálenosti závisí na měřítku veličin. Proto je třeba veličiny normovat. Konkrétní hodnota se obvykle dělí nějakou jinou hodnotou: průměrem, směrodatnou odchylkou nebo rozpětím (max–min). Navíc předpokládáme stejný rozptyl u všech veličin. V případě různého rozptylu se doporučuje použít *Mahalanobisovu vzdálenost*, která je zobecněním vzdálenosti eukleidovské

$$d_{M^2}(\boldsymbol{x}_1, \boldsymbol{x}_2) = (\boldsymbol{x}_1 - \boldsymbol{x}_2)^T \mathbf{S}^{-1} (\boldsymbol{x}_1 - \boldsymbol{x}_2) .$$

Z používaných metod shlukové analýzy se zmiňme o:

- hierarchickém shlukování,
- metodě *K*-středů (*K*-means clustering).

Při hierarchickém shlukování se obvykle postupuje metodou „zdola nahoru". Začíná se tedy v situaci, kdy každý příklad tvoří jeden samostatný shluk. Postupně se pak jednotlivé shluky spojují, až skončíme s jedním shlukem obsahujícím všechny příklady (obr. 30). Vzdálenost mezi shluky lze stanovit různým způsobem:

Algoritmus hierarchického shlukování

Inicializace
1. urči vzájemné vzdálenosti mezi všemi příklady
2. zařaď každý příklad do samostatného shluku

hlavní cyklus
1. dokud je více než jeden shluk
   1.1. najdi dva navzájem nejbližší shluky a spoj je
   1.2. spočítej pro tento nový shluk jeho vzdálenost od ostatních shluků

Obr. 30 Algoritmus hierarchického shlukování.

- *metodou nejbližšího souseda* – vzdálenost mezi shluky $\mathcal{U}$ a $\mathcal{V}$ je dána minimem ze vzdálenosti mezi jejich příklady

$$D(\mathcal{U},\mathcal{V}) = \min_{k,l} d(\boldsymbol{x}_k, \boldsymbol{x}_l), \quad \boldsymbol{x}_k \in \mathcal{U}, \ \boldsymbol{x}_l \in \mathcal{V},$$

- *metodou nejvzdálenějšího souseda* – vzdálenost mezi shluky $\mathcal{U}$ a $\mathcal{V}$ udává maximum ze vzdálenosti mezi jejich příklady

$$D(\mathcal{U},\mathcal{V}) = \max_{k,l} d(\boldsymbol{x}_k, \boldsymbol{x}_l), \quad \boldsymbol{x}_k \in \mathcal{U}, \ \boldsymbol{x}_l \in \mathcal{V},$$

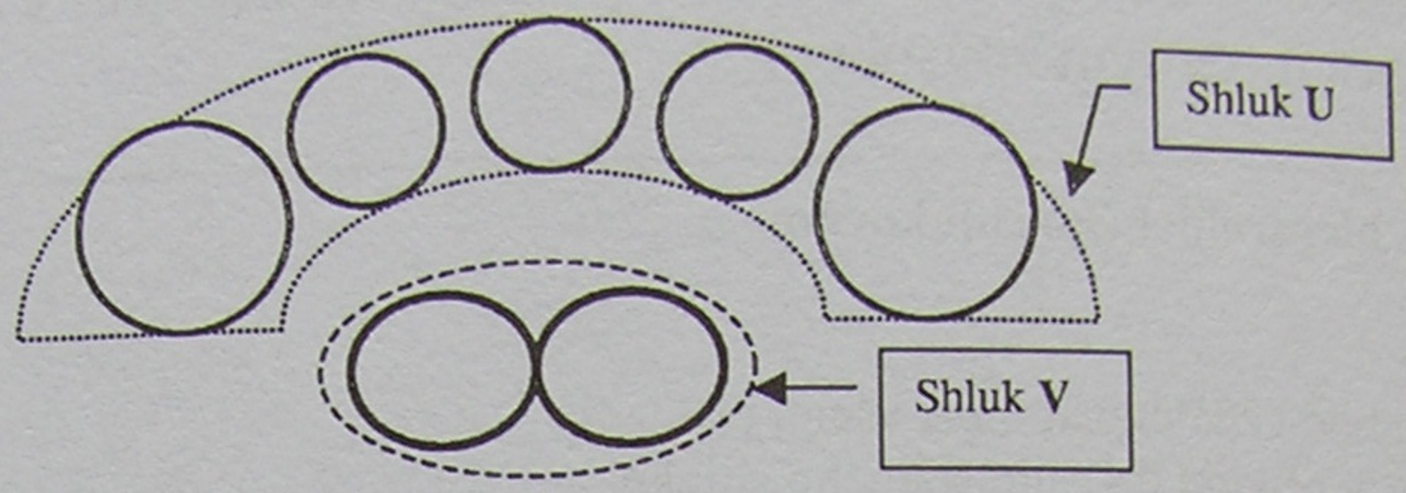
- *metodou průměrné vzdálenosti* – vzdálenost mezi shluky $\mathcal{U}$ a $\mathcal{V}$ je dána průměrem ze vzdálenosti mezi jejich příklady (n_U je počet příkladů ve shluku $\mathcal{U}$ a n_V je počet příkladů ve shluku $\mathcal{V}$)

$$D(\mathcal{U},\mathcal{V}) = \frac{1}{n_U n_V} \sum_{k=1}^{n_U} \sum_{l=1}^{n_V} d(\boldsymbol{x}_k, \boldsymbol{x}_l),$$

- *centroidní metodou* – vzdálenost mezi shluky $\mathcal{U}$ a $\mathcal{V}$ představuje vzdálenost mezi středy shluků

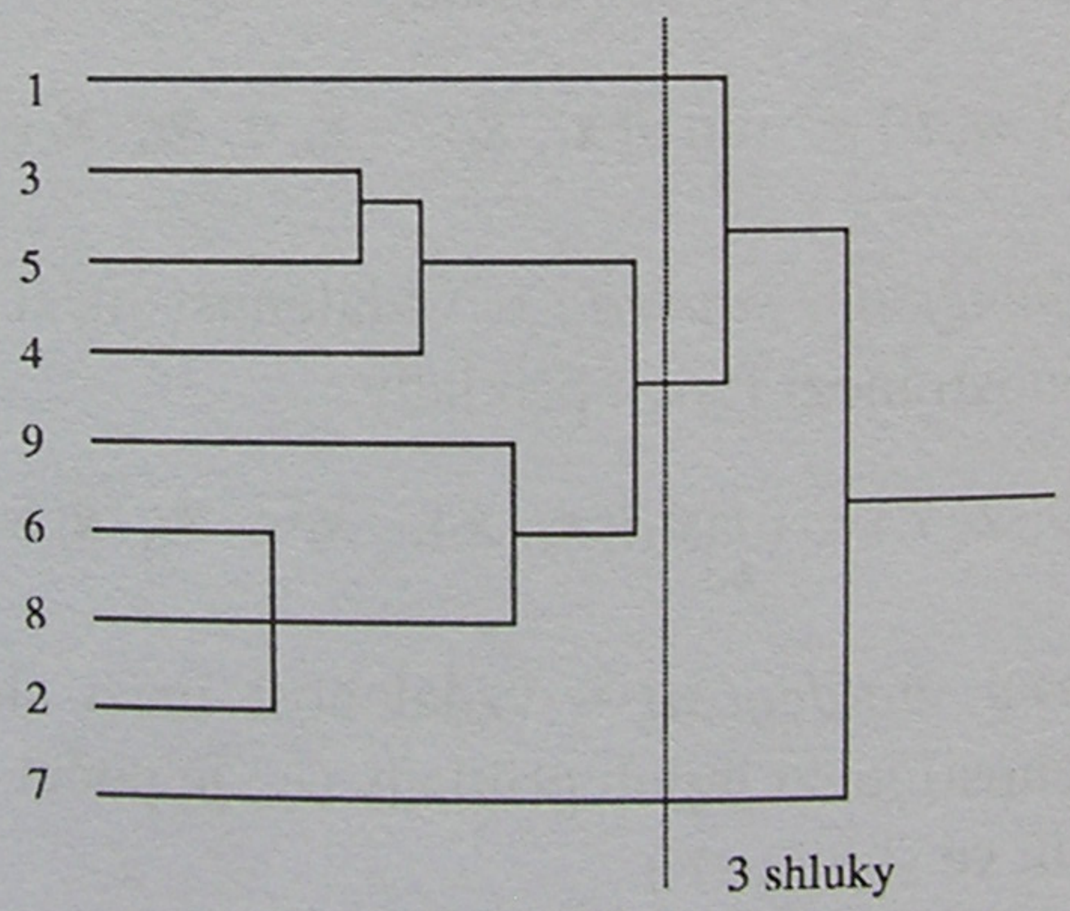
$D(\mathcal{U},\mathcal{V}) = d(\bar{\boldsymbol{u}}, \bar{\boldsymbol{v}})$, $\bar{\boldsymbol{u}}$ je střed shluku $\mathcal{U}$ a $\bar{\boldsymbol{v}}$ je střed shluku $\mathcal{V}$.

Centroidy (středy shluků), zmíněné v předcházejícím výčtu, představují jakési prototypy reprezentující jednotlivé shluky[17]. Nemusí ale platit, že ke každému shluku patří právě jeden centroid. V závislosti na tvaru shluku a zvolené míře pro výpočet vzdálenosti může být jeden shluk reprezentován více centroidy (na obr. 31 znázorněnými jako tučné kružnice)[18].



Obr. 31 Více centroidů pro jeden shluk.

Proces hierarchického shlukování bývá zachycen v podobě tzv. *dendrogramu*. Ten ukazuje (zleva doprava) postupné spojování shluků počínaje očíslovanými příklady. Optimální počet shluků zde není předem znám, odvodíme ho rozborem výsledků – tak, že někde dendrogram „rozřízneme“ (obr. 32).



Obr. 32 Dendrogram.

[17] V nejjednodušším případě můžeme centroidy chápat jako příklady, které nabývají průměrných hodnot jednotlivých veličin v rámci daného shluku. Takto chápané centroidy zaručují optimální klasifikaci v situaci, kdy aposteriorní pravděpodobnosti zařazení příkladů do tříd se řídí normálním rozdělením se stejnou (jednotkovou) kovarianční maticí a stejnou pravděpodobností jednotlivých tříd – podrobnější údaje najde čtenář v oddílu 3.3.

[18] Všimněme si toho, že dané shluky nejsou lineárně separabilní. Lineární diskriminační analýza tedy nedokáže bezchybně od sebe odlišit příklady obou shluků.

Při shlukování metodou *K*-středů předpokládáme, že víme do kolika shluků je možné příklady rozdělit. Počet shluků se tedy během výpočtu nemění, mění se pouze zařazení příkladů k těmto shlukům. Proto je tato metoda méně výpočetně náročná než hierarchické shlukování (a tudíž vhodnější pro větší datové soubory). Příslušný algoritmus je uveden na obr. 33.

Metoda *K*-středů

1. náhodně zvol rozklad do K shluků
2. urči centroidy pro všechny shluky v aktuálním rozkladu
3. pro každý příklad $\boldsymbol{x}$
   3.1. urči vzdálenosti $d(\boldsymbol{x},\boldsymbol{c}_k)$, $k = 1,\ldots,K$, kde $\boldsymbol{c}_k$ je centroid k-tého shluku
   3.2. nechť $d(\boldsymbol{x},\boldsymbol{c}_l) = \min_k d(\boldsymbol{x},\boldsymbol{c}_k)$
   3.3. není-li $\boldsymbol{x}$ součástí shluku l (k jehož centroidu $\boldsymbol{c}_l$ má nejblíže), přesuň $\boldsymbol{x}$ do shluku l
4. došlo-li k nějakému přesunu, potom jdi na 2, jinak konec

Obr. 33 Algoritmus shlukování metodou *K*-středů.

Uvedený algoritmus může mít určité varianty:

- místo počátečního rozkladu lze za centroidy prohlásit prvních K příkladů; odpadne tak krok 2 při prvním průchodu daty,
- přepočet centroidů lze provádět po každém přesunu (tedy v cyklu v kroku 3).

Výsledné shluky jsou při použití metody *K*-středů reprezentovány svými centroidy. Tuto reprezentaci lze snadno použít pro zařazování nových příkladů. Příklad bude (ve shodě s krokem 3.3 algoritmu) zařazen do shluku, k jehož centroidu má nejblíže.

Literatura

Anděl J.: *Matematická statistika*. SNTL/ALFA, Praha/Bratislava 1978.

Havránek T.: *Statistika pro biologické a lékařské vědy*. Academia, Praha 1993.

Hebák P., Hustopecký J.: *Vícerozměrné statistické metody s aplikacemi*. SNTL, Praha 1987.

Kotek Z., Chalupa V., Brůha I., Jelínek J.: *Adaptivní a učící se systémy*. SNTL, Praha 1980.

Michie D., Spiegelhalter D. J., Taylor C.: *Machine learning, neural and statistical classification*. Ellis Horwood, 1994.