Shlukování

přednášky – doplňující poznámky

Jan Outrata

leden 2008

Úvod

- člověk shlukuje "od samotného vývinu mozku (vyšších mozkových funkcí)", shlukování = seskupování podobných objektů (někdy označováno jako **klasifikace**), pojmenovávání shluků
- historie: bilogické klasifikace (Aristoteles – zvířata, Theophrastos – rostliny, Linné – taxonomie rostlin), chemické prvky (Mendělejev), hvězdy, a mnohé další
- jedny objekty lze shlukovat více způsoby (podle různých charakteristik), např. objekty = lidé, charakteristiky = ekonomické ukazatele, rasy, věk, apod.
- použití numerických metod v procedurách shlukování (objektivita, stabilita)
- různá pojmenování shlukování (**shlukové analýzy**) v různých oblastech: numerická taxonomie (biologie), Q-analýza (psychologie), rozpoznávání vzorů bez učitele (informatika, strojové učení), segmentace (ekonomie)

Úvod

- *co je to shluk?* – těžké matematicky definovat, v jeho zobrazení hrají roli vzdálenosti mezi objekty
- **shlukování = hledání shluků** v předem neznámé struktuře objektů, hrozí nebezpečí umělého vnucování struktury!
- **POZOR!** shlukování versus **třídění** = identifikace objektů s předem danou strukturou, "škatulkování"
- typy shlukování (shlukovacích metod):
  - neinkrementální – na vstupu známá (pevně daná) všechna data, nad kterými se jednorázově vytvoří model
    - hierarchické: aglomerativní a divizivní
    - nehierarchické: optimalizační, pravděpodobnostní a statistické, ...
    - jiné klasifikace – crisp vs. fuzzy, s nepřekrývajícími vs. překrývajícími se shluky, ...
  - inkrementální – data (objekty) na vstupu (neustále) přibývají, vytvářený model se upravuje podle nových dat
- aplikace: biologie (botanika), medicína (psychologie, psychiatrie), geografie, astrononie, archeologie, ekonomie, ...

Vstupní data

- **objekty** (záznamy) popsány **atributy** (znaky) různých typů
- atributy mohou mít různou **váhu** (rozměr) $\Longrightarrow$ **standardizace atributů** na bezrozměrné veličiny (může data zkreslovat!)
  - původní hodnoty z_{ij} pro objekt i a atribut j
  - průměrná hodnota $\bar{z}_j$ hodnot pro atribut j:

$$\bar{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij}$$

  - směrodatná odchylka $s_j^{(z)}$ hodnot pro atribut j:

$$s_j^{(z)} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2}$$

  - nové hodnoty x_{ij}:

$$x_{ij} = \frac{z_{ij} - \bar{z}_j}{s_j}$$

  - průměrná hodnota $\bar{x}_j = 0$ a směrodatná odchylka $s_j^{(x)} = 1$

Vstupní data

- objekty mohou mít různou "velikost" $\Longrightarrow$ **normalizace objektů** jako vektorů hodnot atributů (může mít velký vliv na výsledek shlukování!, záleží na použité (ne)podobnosti mezi objekty)
- zobrazení objektů (**geometrický model dat**): objekty jako body v prostoru E^n, projekce do prostoru $E^m, m \ll n$ (nejčastěji roviny, $m = 2$) pomocí **metody hlavních komponent**

Zobrazení shluků

pro počet atributů $n = 1$ nebo $n = 2$:

- **histogram** = sloupcový graf počtu nebo frekvencí objektů na hodnotách atributů (konkrétních nebo z intervalu) – vysoké sloupce naznačují shluky
- **scatterplot** – objekty zobrazeny jako tečky v prostoru vymezeném hodnotami atributů – hustě "vytečkovaná" místa naznačují shluky
- graf odhadu $\hat{f}(x)$ funkce $f(x)$ hustoty pravděpodobnosti hodnot z intervalu $(x - h, x + h)$ (kernel density estimators), viz literatura
- **vrstevnice** – nahuštěné naznačují shluky
- ...

pro $n \geq 3$:

- matice zobrazení pro $n' = 2$ "pro všechny možné podmnožiny atributů velikosti 2"
- projekce prostoru E^n do prostoru $E^m, m \ll n$ (nejčastěji roviny, $m = 2$) pomocí **metody hlavních komponent**