Shluková analýza

Jan Kelbel David Šilhán

Obsah

1 Úvod

Shluková analýza patří mezi metody učení bez učitele. Jejím cílem je v dané množině objektů nalézt její podmnožiny – shluky objektů – tak, aby si členové shluku byli navzájem podobní, ale nebyli si příliš podobní s objekty mimo tento shluk.

1.1 Formulace úlohy

Nechť $\mathcal{X}$ značí množinu n objektů. Rozklad $\Omega = \{C_1, C_2, \ldots, C_m\}$ množiny $\mathcal{X}$ je množina disjunktních, neprázdných podmnožin $\mathcal{X}$, které dohromady tvoří $\mathcal{X}$. To jest, pro $i \neq j$

$$C_i \cap C_j = \emptyset, \qquad C_1 \cup C_2 \cup \cdots \cup C_m = \mathcal{X}.$$

Každá množina C_i se nazývá komponentou rozkladu.

Shlukování je rozklad množiny $\mathcal{X}$. Komponenty tohoto rozkladu se nazývají shluky. Shlukování patří k třídě úloh následujícího tvaru:

Nechť x je náhodný objekt z množiny objektů $\mathcal{X}$ s rozdělením pravděpodobnosti $p_x : \mathcal{X} \to \mathbf{R}$. Nechť D je množina rozhodnutí taková, že pro každý objekt $x \in \mathcal{X}$ se určí jisté rozhodnutí $d \in D$. Nechť $W : \mathcal{X} \times D \to \mathbf{R}$ je pokutová funkce, jejíž hodnota $W(x, d)$ představuje ztráty v případě volby rozhodnutí d pro objekt x. Při zvoleném rozhodnutí d zde pokuta závisí na známém pozorování x a ne na nepozorovatelném stavu. Pojem nepozorovatelný stav se zde vůbec nevyskytuje.

Cílem úlohy je zkonstruovat strategii $Q : \mathcal{X} \to D$, která minimalizuje hodnotu:

$$\sum_{x \in \mathcal{X}} p(x) \cdot W(x, Q(x))$$

1.2 Typy metod shlukové analýzy

Shlukovací metody můžeme rozdělit podle cílů, k nimž směřují, na *hierarchické* a *nehierarchické*.

Mějme množinu objektů $\mathcal{X}$. Hierarchické shlukování je systém navzájem různých neprázdných podmnožin množiny $\mathcal{X}$, v němž průnikem každých dvou podmnožin je buď jedna z nich nebo prázdná množina a v němž existuje alespoň jedna dvojice podmnožin, jejichž průnikem je jedna z nich.

Nehierarchické shlukování je systém navzájem různých neprázdných podmnožin množiny $\mathcal{X}$, v němž průnikem každých dvou podmnožin není žádná z nich. My se budeme věnovat pouze metodám vedoucím k disjunktnímu rozkladu množiny $\mathcal{X}$.

2 Objekty a znaky

Objekty určenými ke klasifikaci jsou předměty nebo jevy. Každý konkrétní objekt je popsán p-ticí stavů předem stanovených *p znaků*. Stavům zpravidla přiřazujeme jejich číselné kódy. Tato čísla pak představují hodnoty znaků. *Objektem* pro shlukovou analýzu je tedy p-rozměrný vektor čísel.

2.1 Typy znaků

Znakem rozumíme zobrazení množiny objektů určených ke klasifikaci do množiny stavů znaku. Znaky objektů a množiny jejich stavů mohou být:

1. Kvalitativní znaky

   - Konečná množina popisujících termínů, kterým mohou být přiřazeny číselné kódy. Rozlišují se nominální, např. barva (1 - červená, 2 - bílá, 3 - žlutá), a ordinální, které se dají uspořádat, např. zeleň listu (1 - světlá, 2 - střední, 3 - tmavá).

   - Binární (*dichotomické*) znaky, např. mít modré oči (pravda/nepravda) nebo pohlaví (muž/žena).

2. Kvantitativní znaky

   - Interval v reálných nebo celých číslech, např. délka, teplota.

2.2 Standardizace dat

Hodnoty jednotlivých znaků objektů jsou často v různých jednotkách. To může způsobovat, že se určité znaky jeví jako dominující a jiné znaky jen málo ovlivňují průběh shlukování. Někdy je proto výhodné data upravit tak, aby byly všechny znaky souměřitelné. Jedním ze způsobů, jak toho docílit, je *standardizace dat.*

Nechť je dána matice dat $\boldsymbol{Z} = (z_{ij})$ typu $n \times p$, jejíž řádky jsou p-rozměrné vektory čísel charakterizující n objektů. Standardizaci dat provedeme ve dvou krocích:

1. Vypočteme střední hodnotu $\bar{z}_j$ j-tého znaku z_j a směrodatnou odchylku s_j pro $j = 1, 2, \ldots p$ podle vzorců:

$$\bar{z}_j = \frac{1}{n} \sum_{i=1}^{n} z_{ij}\,, \qquad s_j = \left[\frac{1}{n} \sum_{i=1}^{n} (z_{ij} - \bar{z}_j)^2 \right]^{1/2}$$

2. Původní hodnoty z_{ij} j-tého znaku i-tého objektu přepočteme na tzv. standardizované hodnoty:

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

Tyto standardizované hodnoty znaků mají nyní střední hodnotu rovnu 0 a rozptyl 1.

2.3 Normalizace objektů

Objekty pro shlukovou analýzu, jak jsme již uvedli, jsou určeny vektory o p složkách představujících hodnoty vybraných p znaků. Normy těchto vektorů mohou někdy nežádoucím způsobem ovlivňovat výsledky kvantitativního hodnocení podobností objektů. V takových případech je vhodné *normalizovat* tyto vektory, aby měli stejnou normu (nejlépe jednotkovou).

3 Podobnost objektů

Jedním ze základních problémů shlukové analýzy je pojetí vzájemné podobnosti objektů a kvantitativní vyjádření této podobnosti. Tedy stanovení vhodného předpisu π přiřazujícího každé dvojici (O_r, O_s) objektů číslo $\pi(O_r, O_s)$, které budeme považovat za míru podobnosti objektů, tak, aby byly splněny alespoň tyto požadavky:

$\pi(O_r, O_s) \geq 0$

$\pi(O_r, O_s) = \pi(O_s, O_r)$

Dále má smysl požadovat, aby pro $O_r = O_s$ nabývala funkce $\pi(O_r, O_s)$ maxima z oboru hodnot π.

Předpis π vyjadřující podobnost objektů dává tím větší hodnotu $\pi(O_r, O_s)$, čím větší je vzájemná podobnost objektů. U většiny shlukovacích metod je ale vhodnější vycházet z míry nepodobnosti objektů. V tom případě nabývá $\pi(O_r, O_s)$ pro $O_r = O_s$ minimální hodnoty, tj. hodnoty nula.

Třemi základními typy předpisu π jsou koeficienty asociace a koeficient korelace, představující míry podobnosti objektů, a metriky představující míry nepodobnosti objektů.

3.1 Koeficienty asociace

Tyto koeficienty jsou určeny pro hodnocení podobnosti objektů popsaných výhradně dichotomickými znaky. Nechť jsou objekty charakterizovány p znaky nabývajícími hodnot z $\{0, 1\}$. Asociaci dvojice objektů (O_r, O_s) určují čísla označená $\alpha_1, \beta, \gamma, \alpha_0$, kde:

α_1 je počet případů pozitivní shody, tj. počet znaků nabývajících hodnoty *pravda* pro oba objekty, tj. $\alpha_1 = \sum_{i=1}^{p} \min\{x_{ri}, x_{si}\}$.

β je počet znaků takových, že pro objekt (O_r) mají hodnotu *pravda* a pro (O_s) hodnotu *nepravda*, tj. $\beta = \sum_{i=1}^{p} x_{ri} - \alpha_1$.

γ je počet znaků takových, že pro objekt (O_r) mají hodnotu *nepravda* a pro (O_s) hodnotu *pravda*, tj. $\gamma = \sum_{i=1}^{p} x_{si} - \alpha_1$.

α_0 je počet případů negativní shody, tj. $\alpha_0 = p - (\alpha_1 + \beta + \gamma)$.

Z těchto hodnot vycházejí koeficienty asociace. Většina těchto koeficientů je definována tak, aby oborem jejich hodnot byl interval $\langle 0, 1\rangle$.

3.2 Metriky

Jeden z nejběžnějších způsobů vyjádření podobnostních vztahů mezi objekty jsou metriky vycházející z geometrického modelu dat. Přiřadíme-li objektům charakterizovaným p znaky jako modely body p-rozměrného euklidovského prostoru $\mathbf{E}_p$, pak je pro dva body (r, s) definována euklidovská vzdálenost:

$$\delta(r, s) = \left[\sum_{i=1}^{p} (x_{ri} - x_{si})^2\right]^{1/2}$$

Obecně je *metrika* δ funkce definovaná na $\mathbf{E}_p \times \mathbf{E}_p$ přiřazující každé dvojici bodů (r, s) číslo $\delta(r, s)$ splňující tyto čtyři podmínky:

$\delta(r, s) = 0 \iff r = s$ (identita)
$\delta(r, s) \geq 0$
$\delta(r, s) = \delta(s, r)$ (symetrie)
$\delta(r, t) \leq \delta(r, s) + \delta(s, t)$ (trojúhelníková nerovnost)

Minkowského metrika je definována vztahem

$$\delta_k(r, s) = \left(\sum_{i=1}^{p} |x_{ri} - x_{si}|^k\right)^{1/k}$$

Pro různá m dostáváme metriky:
$m = 1$ vzdálenost v městských blocích
$m = 2$ euklidovská metrika
$m \to \infty$ sup–metrika $\delta_\infty(r, s) = \max\limits_{i=1,\ldots p} \{|x_{ri} - x_{si}|\}$

3.3 Koeficienty nepodobnosti objektů

Koeficienty nepodobnosti objektů $\boldsymbol{d}$ jsou jednotící předpis pro hodnocení podobnostních vztahů objektů. Na základě koeficientů uvedených v 3.1 a metrik 3.2 je můžeme definovat pro objekty O_r a O_s takto:

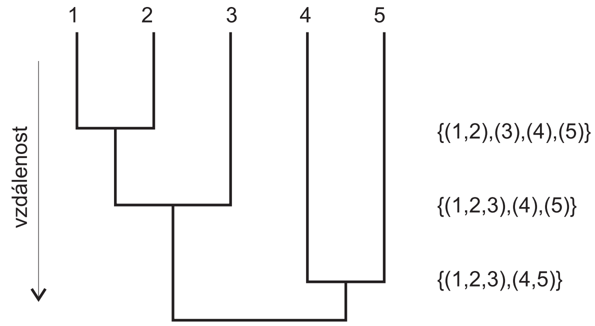
- Pro koeficienty asociace označené S je koeficient nepodobnosti objektů $\boldsymbol{d}(O_r, O_s) = 1 - S$.
- Pro metriky $\delta(X, Y)$ je $\boldsymbol{d}(X, Y) = \delta(X, Y)$.

4 Hierarchické shlukování

Hierarchické shlukování je sekvence vnořených rozkladů, která na jedné straně začíná triviálním rozkladem, kdy každý objekt dané množiny objektů tvoří jednoprvkový shluk, a na druhé straně končí triviálním rozkladem s jedním shlukem obsahujícím všechny objekty.

Podle směru postupu při shlukování dělíme metody hierarchického shlukování na aglomerativní a divizivní.

Dendrogram je binární strom znázorňující hierarchické shlukování. Každý uzel tohoto stromu představuje shluk. Horizontální řezy dendrogramem jsou rozklady ze shlukovací sekvence. Vertikální směr v dendrogramu představuje „vzdálenost“ mezi shluky (rozklady).



Obrázek 1: Dendrogram pro pět objektů

4.1 Aglomerativní hierarchické shlukování

Počáteční rozklad množiny objektů $\mathcal{X}$ tvořený n jednoprvkovými shluky označíme jako nultý rozklad Ω_0. Abychom mohli postupně vytvářet další rozklady množiny objektů, definujeme způsob hodnocení podobnostních vztahů mezi shluky. V každém kroku s shlukování pak vybíráme ty dva shluky, které jsou si ve smyslu naší definice nejpodobnější. Tyto dva shluky sloučíme a vytvoříme tak nový shluk. Vznikne rozklad Ω_s. Každý z vytvořených rozkladů sníží počet shluků o jeden. Hierarchickým shlukováním je pak posloupnost $n-1$ rozkladů $\Omega_0, \ldots, \Omega_{n-1}$. Rozklad Ω_s je zjemněním rozkladu Ω_{s+1}.

Rekurzivní definice

1. Nultý rozklad množiny objektů $\Omega_0 = \{C_{01}, C_{02}, \ldots, C_{0n}\}$, kde $C_{0j} = \{O_j\}$.

2. V i-tém kroku $(0 < i < n-1)$ najdeme jedinou dvojici shluků (C_{ir}, C_{is}), pro kterou je hodnota koeficientu nepodobnosti shluků $\boldsymbol{D}$ (viz 4.2) minimální

$$\boldsymbol{D}(C_{ir}, C_{is}) = \min\{\boldsymbol{D}(C_{iu}, C_{iv})\}$$

   kde minimum se bere přes všechny shluky rozkladu Ω_i. Sjednocení $C_{ir} \cup C_{is}$ nahradí v rozkladu Ω_{i+1} samostatné shluky C_{ir}, C_{is}.

3. V posledním $(n-1)$ kroku jsou všechny shluky sjednoceny v jediný shluk. $\Omega_{n-1} = \mathcal{X}$.

4.2 Koeficienty nepodobnosti shluků

Nechť $\boldsymbol{d}$ je libovolný koeficient nepodobnosti objektů, C_i, C_j jsou shluky rozkladu. $\boldsymbol{D}(C_i, C_j)$ je koeficient nepodobnosti shluků. Existuje více metod jeho vyjádření.

Metoda nejbližšího souseda, známá také pod názvem „single link", definuje koeficient nepodobnosti shluků vztahem

$$\boldsymbol{D}(C_i, C_j) = \min_{\substack{O_i \in C_i \\ O_j \in C_j}} \{\boldsymbol{d}(O_i, O_j)\}$$

Metoda nejvzdálenějšího souseda, jinak také „complete link", je definována vztahem

$$\boldsymbol{D}(C_i, C_j) = \max_{\substack{O_i \in C_i \\ O_j \in C_j}} \{\boldsymbol{d}(O_i, O_j)\} \qquad \text{pro } C_i \neq C_j$$

$$D(C_i, C_i) = 0$$

Dalšími metodami pro určení nepodobnosti shluků jsou *centroidní metoda*, *mediánová metoda* a jiné (viz např. [1]).

Při hierarchickém shlukování můžeme počítat nepodobnost nově vzniklého shluku s ostatními na základě uchovávaných hodnot koeficientů nepodobnosti už existujících shluků. Koeficient nepodobnosti $\boldsymbol{D}(C_r, C_s)$ shluku C_r se shlukem $C_s = C_p \cup C_q$ se vypočte podle vztahu

$$\boldsymbol{D}(C_r, C_s) = \alpha_i \boldsymbol{D}(C_r, C_p) + \alpha_j \boldsymbol{D}(C_r, C_q) + \beta \boldsymbol{D}(C_p, C_q) + \gamma \left|\boldsymbol{D}(C_r, C_p) - \boldsymbol{D}(C_r, C_q)\right|.$$

Koeficienty $\alpha_i, \alpha_j, \beta$ a γ se mění v závislosti na použité metodě, jakou je definována nepodobnost shluků.

4.3 Divizivní hierarchické shlukování

Na rozdíl od aglomerativních metod vytvářejí divizivní metody hierarchický systém rozkladů množiny objektů postupným rozdělováním existujících shluků. Při aplikaci divizivního algoritmu postupujeme tak, že za počáteční shluk uvažujeme celou množinu objektů a postupně rozdělujeme existující shluky, až jsou všechny shluky jednoprvkové.

Postup divizivního shlukování spočívá v postupném rozdělování každého z existujících shluků na dva nové tak, aby výsledný rozklad tohoto shluku byl optimální vzhledem k nějakému kritériu. Nalezení absolutně optimálního rozkladu množiny n objektů na dvě podmnožiny však vyžaduje prozkoumání $2^{n-1} - 1$ možností. Tento postup je prakticky proveditelný jen pro malý počet objektů.

MacNaughton–Smithova metoda [1] je aplikovatelná i na rozsáhlejší množiny objektů při nevelkých nárocích na čas počítače.

5 Nehierarchické shlukování

5.1 Nalezení optimálního počtu shluků

Pro nalezení optimálního počtu shluků před samotným hledáním rozkladu je vhodné použít jednoho z indexu jehož optimální hodnotu lze nalézt v závisloti na proměnném parametru K.

- **Calinski-Harabascův index** (nejlepší výsledky z 30 kritérií testovaných Milliganem a Cooperem), pro optimální K se hledá maximum CH(K)

$$CH(K) = \frac{n-K}{K-1}\left[\frac{E_1^2}{E_K^2} - 1\right]$$

  n je počet obrazů, K počet shluků a E_K^2 je kvadrát příslušného funkcionálu

- **C index** je normalizovaný tvar Γ statistiky

$$\Gamma = \sum_{q=1}^{n-1} \sum_{r=q+1}^{n} d(q,r)c(q,r)$$

  Kde $c(q,r) = 1$ pokud jsou x_q a x_r ve stejném shluku, v opačném případě $c(q,r) = 0$. $d(q,r)$ je vzájemná nepodobnost nebo Euklidovská vzdálenost mezi dvěma vzory.

$$C(K) = \frac{\Gamma - \min(\Gamma)}{\max(\Gamma) - \min(\Gamma)}$$

  kde $\min(\Gamma)$ je suma a_k nejmenších nepodobností

  $\max(\Gamma)$ je suma a_k největších nepodobností

- **Goodman-Kruskal**, pro optimální K se hledá maximum $\gamma(K)$,

$$\gamma(K) = \frac{S(+) - S(-)}{S(+) + S(-)}$$

  kde $S(+)$ znamená počet souhlasných a $S(-)$ počet nesouhlasných čtveřic čísel $[d(q,r), d(s,t), f(q,r), f(s,t)]$, kde čtvěřice je souhlasná pokud $d(q,r) < d(s,t)$ a $f(q,r) < f(s,t)$ nebo $d(q,r) < d(s,t)$ a $f(q,r) < f(s,t)$. $d(q,r)$ je Euklidovská vzdálenost a funkce $f(q,r) = 1 - c(q,r)$ je rovna $f(q,r) = 1$ pokud x_q a x_r jsou ve stejném shluku a rovna $f(q,r) = 0$ pokud x_q a x_r jsou v jiném shluku.

5.2 Problém počátečního rozkladu

Stanovení optimálního počátečního rozkladu na k-shluků při počátku rozkladu může být následováno buď zachováním stejného počtu shluků nebo změnou počtu k v průběhu výpočtu v závislosti na řídících proměnných algoritmu. Zvláště v prvním typu algoritmů, kdy už nedochází ke změnám v počtu shluků k je nutno klást na počáteční rozklad důraz. Počet tříd může být stanoven některou z dříve uvedených metod. Dalším krokem klasifikace je stanovení typických vzorových objektů (v Euklidovském prostoru typických bodů) kolem nichž se dá předpokládat vytvoření shluků. Počáteční rozklad pak bývá z počátečních údajů odvozen tak, aby hodnota funkcionálu kvality rozkladu byla už v počátečním stádiu shlukování co nejlepší. Pro výběr výchozích bodů se může použít jednoho z jednoduchých postupů:

- vybrat prvních k-bodů z libovolně uspořádané množiny bodů
- generovat k umělých bodů, jejichž každá souřadnice je náhodné číslo z příslušného variačního intervalu

- generovat $(2^p + 1)$ počátečních typických bodů (p je počet rozměrů prostoru)z nichž první je těžištěm celé množiny bodů a další pak jsou vrcholy p-rozměrného kvádru se středem v těžišti a hranami délky $2s_i$. (s_i je směrodatná odchylka hodnot i-té souřadnice) rovnoběžnými s osami soustavy souřadnic. Můsí platit $2^p > n$, kde n je počet bodů.

Obě metody jsou založeny na střídání dvou kroků

- výpočet typických bodů existujících skupin shlukových bodů
- sestrojení skupin shlukových bodů přiřazením každého bodu k tomu z existujících typických bodů, k němuž má uvažovaný bod nejblíže

Iterace provádíme do doby dokud nedojdeme ke stabilní konfiguraci, rozkladu na výsledné shluky. Forgyova a Janceyova metoda se liší pouze způsobem výpočtu nových typických bodů vytvořených skupin shlukovacích bodů. Vycházíme-li z počátečního rozkladu na k-skupin, vypočteme počáteční typické body jako těžiště těchto skupin. Dále jednotlivé body přiřazujeme nejbližšímu typickému bodu. Při výpočtu nových typických bodů jednotlivých skupin počítá Forgayova metoda tento typický bod jako těžiště skupiny, zatímco Janceyova metoda umisťuje nový typický bod skupiny do bodu souměrně sdruženého s předcházejícím typickým bodem podle těžiště skupiny. Obě metody lokálně minimalizují funkcionál E_K.

5.3 MacQueenova (K-Means) a Whishartova metoda

MacQueen tento algoritmus publikoval v roce 1967. Algoritmus K-means iterativně hledá hodnoty vektorů tak, že minimalizuje střední odchylku mezi zadanou množinou dat a vektory (vzdálenost bodu od etalonu shluku), které mají k těmto datům nejmenší euklidovskou vzdálenost a rozděluje je do předem daného počtu shluků (tříd) K: $C_1, C_2, \ldots, C_k$.

Základní algoritmus používá Euklidovskou vzdálenost a μ_j je aritmetický průměr bodu ve shluku (střední hodnota ve třídě) – etalon. (zjednodušený ML odhad, kde se nepracuje s pravděpodobnosmi, ale se vzdáleností)

Vstupem algoritmu je množina dat $x_1, x_2, \ldots, x_l$ a číslo K udávající počet vektorů $\mu_j, j = 1, \ldots, k$. Na začátku se inicializují vektory $\mu_j, j = 1, \ldots, k$, na náhodně zvolenou hodnotu nebo použitím nějaké vhodně zvolené heuristiky (např. využívající apriorní znalost o úloze). Po inicializaci se začnou iterativně opakovat následující dva kroky:

1. ***Klasifikace***: Všechna data $x_i, i = 1, \ldots, l$, se klasifikují do tříd určených vektory $\mu_i, i = 1, \ldots, k$, podle minima euklidovské vzdálenosti. Tedy vzor x_i je přiřazen do třídy y_i podle $y_i = \operatorname*{argmin}_j \parallel x_i - \mu_j \parallel$.

2. ***Přepočítání vektorů*** μ_j: Vypočítají se nové hodnoty vektorů μ_j jako střední hodnoty dat x_i, které byly klasifikovány do třídy určené příslušným vektorem μ_j. Tedy nová hodnota μ_j se spočte podle vztahu $\mu_j = \frac{1}{l_j} \sum_{i=1, y_i=j}^{l} (x_i)$, kde l_j je počet vzorů x_i klasifikovaných v druhém kroku do třídy určené vektorem μ_j.

Kroky 1 a 2 se opakují do té doby dokud se alespoň jeden vektor x_i klasifikuje do jiné třídy než byl klasifikován v předcházejícím kroku.

MacQueenova metoda se od Wishartovy liší tím, že nekončí hned po první iteraci při níž nedošlo ke změně v rozkladu do shluků, ale až v následující neměnné iteraci.

Vlastnosti k-means:

1. jednoduchý
2. prvky se mohou přeskupovat mezi shluky
3. pouze pro metrická data
4. konverguje v konečném počtu kroků k nějakému řešení C^*
5. může existovat více řešení C^* v závislosti na počátečních podmínkách

Experimentální zkušenosti naznačují, že je tento algoritmus pro účely vektorové kvantizace velmi vhodný jak z hlediska kvality rozkladu trénovací množiny do shluků, tak i poměrně rozumnými výpočetními nároky.

5.4 Konvergence Forgyovy, Janceyovy a MacQueenovy metody

Z vlastnosti těžiště $T(A_h)$ podmnožiny $A_h = \{O_{h1}, O_{h2}, \dots, O_{hr}\}$ množiny objektů vyplývá, že

$$E_h = \sum_{i=1}^{r} d_E^2(O_{hi}, Y_h)$$

je minimální, platí-li pro pro typický objekt množiny A_h

$$Y_h = T(A_h)$$

Proto je též

$$E = \sum_{j=1}^{k} E_j$$

minimální, platí-li pro všechny podmnožiny $A_j (j = 1, 2, \dots, k)$ rozkladu Ω rovnost $Y_j = T(A_j)$

Vezměme v úvahu dvě skutečnosti. Za prvé je počet různých způsobů, jak lze n objektů rozložit do k podmnožin, konečný a dá se vypočítat jako Stirlingovo číslo druhého druhu

$$S_n^{(k)} = \frac{1}{k!} \sum_{i=0}^{k} (-1)^{k-i} \binom{k}{i}$$

Forgyova metoda střídá dva kroky a to vytvoření podmnožin přiřazením objektů k nejbližšímu typickému objektu (těžišti) a výpočet nových typických objektů (těžišť) vzniklých podmnožin. Je zřejmé, že první z uvedených kroků snižuje celkový součet čtverců chyb E až do docílení rozkladu, kdy součet čtverců chyb zůstává konstantí. Rovněž druhý krok vede přemístěním typických objektů do těžišť vzniklých skupin ke snížení hodnoty E. Proto se žádný rozklad v posloupnosti $\Omega_1, \Omega_2, \dots$ nemůže zopakovat. Z uvedených dvou skutečností vyplývá, že Forgyova metoda je konvergentní. Konvergentní varianta MacQuinovy k-průměrové metody rovněž snižuje v každém kroku hodnotu funkcionálu E, protože i v případě této metody se objekt přemístí z jedné podmnožiny do druhé pouze tehdy, je-li blíže typickému objektu druhé podmnožiny než typickému objektu první podmnožiny. V případě Janceyovy metody je problém konvergence poněkud složitější, protože nelze ani dokázat konvergenci, ani se dosud nepodařilo najít protipříklad.

5.5 Metody měnící počet shluků

Hledání optimálního počtu shluků lze provést i jako součást hledání optimálního rozkladu. Pro všechny takové metody je charakteristické, že kromě udání počátečního počtu k skupin vyžadují zavedení dalších řídících parametrů umožňujících rozhodnutí o tom, zdali má nebo nemá být provedeno sloučení nebo rozdělení skupiny. Mezi metody s řídícími parametry zadanými analytikem které zůstávají stabilní během celého výpočtu patří:

MacQueenova metoda se dvěma parametry (na začátku zadáme počet typických bodů k, slučovací parametr C a rozdělovací parametr R0, prvních k bodů dosadíme za počáteční typické body)

Wishartova metoda RELOC (vyžaduje zadání čtyř řídících parametrů: vzdálenostní práh, minimální počet objektů ve shluku, maximální počet iterací a minimální počet shluků)

Metoda ISODATA (Itertative Self-Organizing Data Analysis Techniques, Ball a Hall. Algoritmus ISODATA pracuje se dvěma typy parametrů. První typ parametrů je zadáván analytikem na počátku výpočtu a po celou dobu zůstávají konstantní (počáteční kulovitost (sphericity) obvykle se volí 1.25, počet iterací, počet dělení v jedné itetraci, očekávaný počet shluků, minimální akceptovatelná změna hodnoty funkcionálu $\sum E_K^2$ mezi dvěma za sebou následujícími iteracemi, maximální počet změn kulovitosti), druhý typ proměnných se mění v čase během výpočtu)

Jiné metody počítají řídící parametry z analyzovaných dat. Představitelem tohoto směru je metoda CLASS Fromma a Notrthouse. Metoda vychází z algoritmu ISODATA a navíc vyžaduje zadání jen 3 vstupních parametrů (maximální počet iterací γ, minimální počet bodů θ ve skupině a rozdělovací práh S_0). Na počátku algoritmus CLASS generuje $2^S + 1$ počátečních typických bodů. Iterace algoritmu CLASS se skládá z následujících kroků:

Vyloučení malých skupin Všechny skupiny o méně než θ bodech, které se nezměnily během posledních dvou iterací, jsou z analýzy vyloučeny.

Rozdělení skupin V m-té iteraci vypočteme rozdělovací práh

$$S_m = S_{m-1} + \frac{1 - S_0}{\gamma}$$

Ve skupině vypočteme nové souřadnice každého bodu jako odchylky od souřadnic těžiště (typického bodu) skupiny. Pro j-tou souřadnici pak vypočteme průměrnou odchylku D_{j1}, resp. D_{j2} bodů ležících vpravo, resp. vlevo od těžiště.

$$D_{j1} = \frac{1}{k_1} \sum_{i=1}^{k_1} x_{ij} \qquad - D_{j2} = \frac{1}{k_2} \sum_{i=1}^{k_2} x_{ij}$$

kde k_1, resp. k_2 je počet bodů vpravo, resp. vlevo od těžiště. V rámci skupiny dále vypočteme parametry a_1, resp. a_2 pro body vpravo, resp. vlevo od těžiště.

$$a_1 = \max_j \frac{D_{j_1}}{\max x_{ij}} \qquad a_2 = \max_j \frac{D_{j_2}}{\min x_{ij}}$$

kde $j = 1, 2, ..., p$ a i probíhá všechny indexy bodů skupiny. Je-li v m-té iteraci počet skupin menší než $2K$,$a_1 > S_m$ nebo $a_2 > S_m$ a zároveň počet bodů větší než $2\theta + 1$, rozdělíme skupinu podle j-té souřadnice. v níž a_1 nebo a_2 nabývá maximální hodnoty a to na část vlevo a část vpravo od j-té souřadnice těžiště skupiny..

Zrušení skupiny Vypočteme průměrnou mininální vzdálenost skupin

$$\tau = \frac{1}{h} \sum_{i=1}^{h} D_i$$

kde h je současný počet skupin, D_i je minimální vzdálenost i-té skupiny od ostatních $h - 1$ skupin vyjádřená vzdáleností těžišť skupin. Platí-li pro i-tou skupinu $D_i < \tau$ a je-li zárovaň počet skupin větší než $K/2$, zrušíme i-tou skupinu a každý její bod přiřadíme ke skupině s nejnižším typickým bodem.

Reference

[1] Alena Lukasová, Jana Šarmanová: *Metody shlukové analýzy.* SNTL, Praha 1985.

[2] C.H. Chen and L.F. Pau. *Handbook of Pattern Recognition and Computer Vision.* World Scientific, 1996.

[3] Michail I. Schlesinger, Václav Hlaváč: *Deset přednášek z teorie statistického a strukturního rozpoznávání.* Vydavatelství ČVUT, Praha 1999.

[4] Josef Psutka: *Komunikace s počítačem mluvenou řečí.* Academia Praha 1995, str. 140

[5] Vladimír Mařík, Olga Štěpánková, Jiří Lažanský: *Umělá inteligence I.* Academia, Praha 1993.

[6] Vladimír Mařík, Zdeněk Kotek a kol.: *Metody rozpoznávání a jejich aplikace.* Academia, Praha 1993.

[7] Petr Jurčík, Vít Vrba: *Shluková analýza,zápis z přednášky.*-, 2001.

[8] Zdeněk Kotek, Petr Vysoký, Zdeněk Zdráhal: *Kybernetika.* SNTL, Praha 1990.