Shluková analýza vícerozměrných dat v programu R

- příklad použití metod PAM, CLARA a fuzzy shlukové analýzy

http://data.tulipany.cz

Úvodní poznámky a popis dat

Pro analýzu vícerozměrných dat existují efektivní algoritmy, často dobře dostupné v softwarových nástrojích pro analýzu dat a dobře popsané v původní literatuře. Z tohoto pohledu je pro úspěšnost reálné analýzy určující dobré porozumění řešené problematice, kvalitní příprava dat a interpretace výstupů výpočtů. Vedle toho je také jistě prospěšné mít přehled o možnostech a principech metod, které je možné použít.

Provedeme analýzu některými moderními[1], v programu R dostupnými algoritmy, zaměřenými především na shlukovou analýzu, které zatím v komerčních softwarových nástrojích nejsou tolik rozšířené. V souvislosti s nižší rozšířeností implementací těchto algoritmů se samozřejmě nabízí otázka, proč tomu tak je. V každém případě tyto novější algoritmy řeší některé problémy metod ostatních. Přitom je možné, že pro mnohé praktické aplikace jsou snad méně sofistikované a méně technicky pokročilé metody postačující.

Budeme pracovat se souborem uměle vytvořených dat, která mají obecně tu výhodu, že můžeme mít dobrou představu o tom, z jakého rozdělení data pocházejí. Generovaný soubor obsahuje z důvodů snahy o dosažení lepší přehlednosti při zobrazení zdrojových dat (v rovině) pouze dvě proměnné[2].

Obvykle není příliš smysluplné dlouze uvažovat o kvalitě modelu vytvořeného nad daty, která vznikla způsobem naznačeným výše[3]. Model, který není návodem k jednání, nemá prakticky žádnou cenu a není zajímavé sledovat, jakých hodnot jednotlivých hodnoticích kritérií takový model dosahuje. Nicméně v případě reálných analýz mohou podobné indikátory kvality, byť jejich konstrukce pro modely zaměřené na nalezení shluků v datech (a snad i struktury vlastní objektům reálného světa) je vděčným předmětem diskusí, poskytnout odhad užitečnosti modelu při jeho případné aplikaci a umožňují tak vybrat model nejvhodnější. Systém R nabízí pro tento účel zajímavý a zatím poměrně málo rozšířený výstup, totiž graf obrysů shluků (Silhouette plot) a charakteristiky s ním spojené.

Nejprve vytvoříme datový soubor. Pro vytvoření datového souboru využijeme generátor pseudonáhodných čísel z vícerozměrného normálního rozdělení, který je v R dostupný v knihovně MASS. Použijeme data ze dvou zdrojů a pro další analýzu je spojíme do souboru se dvěma spojitými proměnnými a 250 případy.

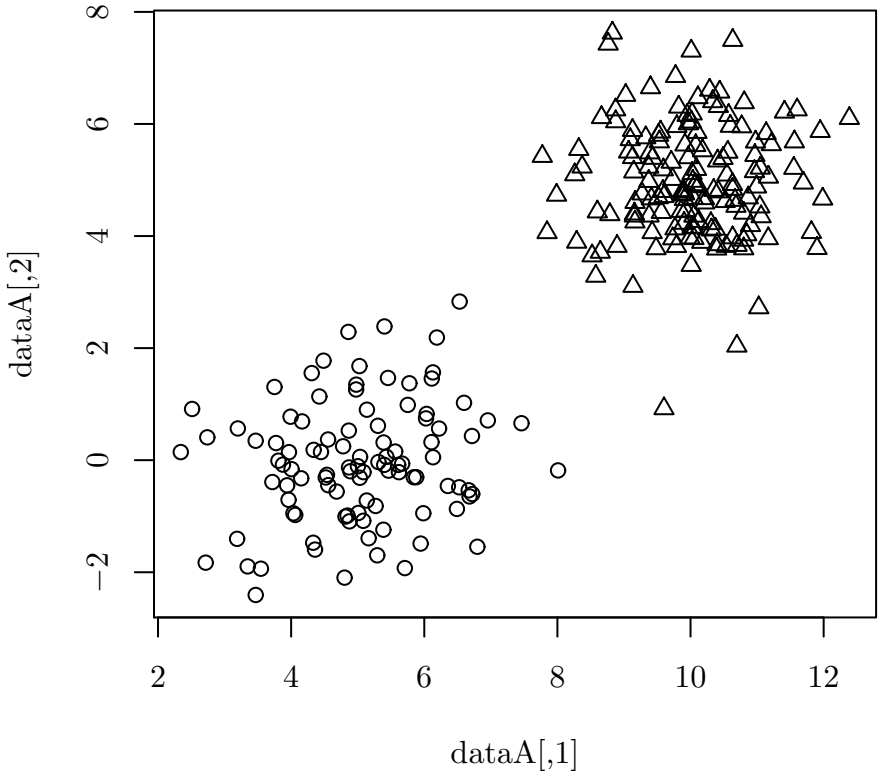
```
> data1 <- mvrnorm(n=150, mu=c(10,5), Sigma=diag(2))
> data2 <- mvrnorm(n=100, mu=c(5,0), Sigma=diag(2))
```

Jak je asi nejlépe patrné z obrázku (1), ve kterém jsou případy z prvního zdroje značeny trojúhelníčky a případy z druhého zdroje kolečky, budeme při shlukování očekávat rozdělení do dvou shluků.

[1] pokud je uvedeno spojení *moderní algoritmy*, může to znamenat algoritmy třeba dvě desítky let staré

[2] je však potřeba připomenout, že systém R standardně podporuje často používané zobrazení shluků v rovině určené dvěma nejdůležitějšími hlavními komponentami, resp. pro vykreslení grafu využívá metodu vícerozměrného škálování při práci s kategoriálními daty

[3] ale někdy ano - například pokud potřebujeme testovat novou metodu a předpokládáme, že bude úspěšná při analýze reálných dat podobně jako při analýze dat umělých, přitom úspěšnost můžeme snadno vyhodnotit spíše nad daty, která již dobře známe



obr. (1)

Značení: v dalším textu označuje k počet shluků a n počet případů v datovém souboru.

Fuzzy shluková analýza

Základní pojmy a principy fuzzy shlukové analýzy v podobě, v jaké bude provedena, jsou již dostatečně zdokumentovány v původní literatuře. Dovolíme si však podle [2] a [1] připomenout některé z pojmů, které se mohou zobrazovat například ve výstupu zpracování dat v R.

Při fuzzy shlukování nejsou případy nutně jednoznačně přiřazeny k určitému shluku, pracuje se s koeficienty příslušnosti i-tého případu k v-tému shluku u_{iv} takovými, že $u_{iv} \in \langle 0, 1\rangle$ a $\sum_{v=1}^{k} u_{iv} = 1$ pro $i = \{1, \ldots, n\}$ a $v = \{1, \ldots, k\}$. Koeficienty příslušnosti jsou určovány iterativním výpočtem tak, aby bylo dosaženo při daném počtu shluků co nejmenší hodnoty účelové funkce

$$\sum_{v=1}^{k} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} u_{iv}^{r} u_{jv}^{r} d(i,j)}{2 \sum_{j=1}^{n} u_{jv}^{r}}, \tag{1}$$

kde $d(i, j)$ je nepodobnost případů i a j a r je koeficient větší než 1 zadaný uživatelem (nebo je přednastaven na hodnotu 2), který v případě, že je bližší hodnotě 1, podporuje řešení spíše podobné jednoznačnému přiřazení případů ke shlukům. Celkovou míru ostrosti rozdělení případů do shluků vyjadřuje *Dunnův koeficient rozdělení*

$$F(k) = \frac{\sum_{i=1}^{n} \sum_{v=1}^{k} u_{iv}^{2}}{n},$$

$F(k) \in \langle 1/k, 1\rangle$. Případně se pracuje s normalizovaným Dunnovým koeficientem rozdělení $(F(k) - 1/k)/(1 - 1/k)$ s hodnotami z intervalu $\langle 0, 1\rangle$. Čím vyšší hodnoty dosáhne Dunnův koeficient, tím více je rozdělení případů do shluků blízké ostrému přiřazení.

Pro některé aplikace může být potřebné jednoznačné přiřazení případů ke shlukům. Potom je i při fuzzy shlukové analýze možno vybrat ke každému případu shluk, pro který má tento případ nejvyšší hodnotu koeficientu příslušnosti, tedy pracovat s tzv. nejbližším ostrým přiřazením (closest hard clustering), podobně jako například při práci s výsledky shlukové analýzy, která pracuje s pravděpodobnostmi, že určitý případ je z určité komponenty směsi.

Nad ostrým rozdělením případů do shluků je v systému R možné vytvořit graf obrysů shluků (Silhouette plot), který nabízí možnost posoudit kvalitu výsledného shlukování a může pomoci při volbě vhodného modelu. Pro každý případ i je definována průměrná nepodobnost objektu i a všech ostatních objektů zařazených do stejného shluku jako i (tento shluk označíme A, počet případů zařazených do tohoto shluku označíme $|A|$):

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i, j).$$

Dále pro každý shluk C odlišný od A se zjišťuje průměrná nepodobnost případu i a shluku C

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i, j).$$

Označíme

$$b(i) = \min_{C \neq A} d(i, C).$$

A definujeme *šířku obrysu* případu i jako

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}.$$

Zřejmě jde o číslo z intervalu $\langle -1, 1\rangle$ a snese následující interpretaci:

$s(i) \approx 1$	$\rightarrow$	případ je do shluku dobře zařazen,
$s(i) \approx 0$	$\rightarrow$	případ leží na rozhraní shluků,
$s(i) < 0$	$\rightarrow$	případ je nejspíše zařazen do neodpovídajícího shluku.

Z hodnot šířky obrysu pro jednotlivé případy je možné vypočítat průměrné šířky obrysu pro jednotlivé shluky a podobně celkovou (případy váženou) průměrnou šířku obrysu. Ta umožňuje posoudit kvalitu nalezené klasifikace - v [1] je uvedena doporučená interpretace, podle které hodnota mezi 0,71 a 1 vypovídá o nalezení silné klasifikační struktury, hodnota v rozmezí 0,7 a 0,51 o nalezení přijatelné struktury, hodnota mezi 0,26 a 0,5 naznačuje slabší a možná umělé vztahy a hodnota nižší znamená, že žádná výrazná klasifikační struktura nalezena nebyla.

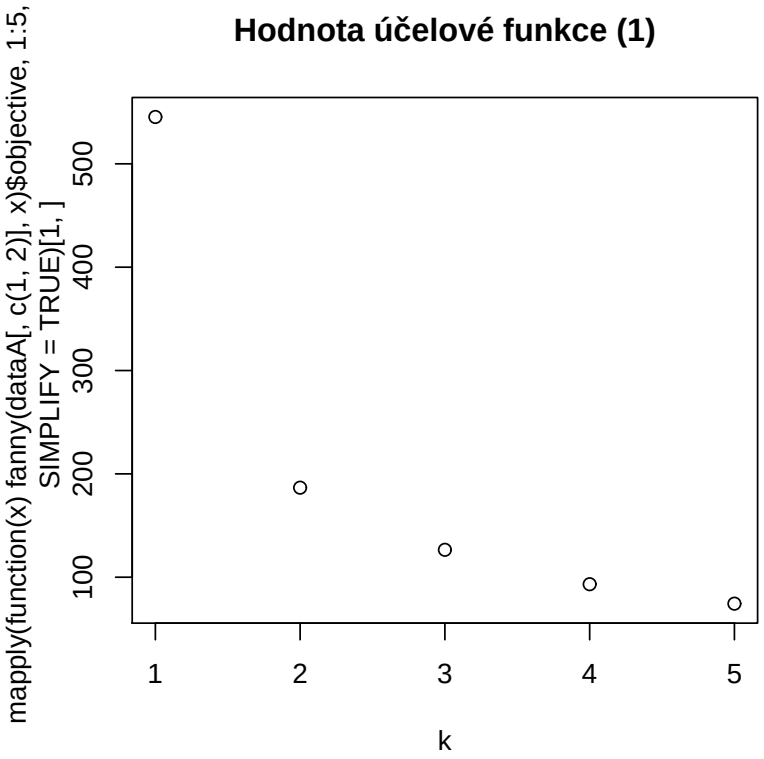
Při grafickém zobrazení obrysů jsou tyto seřazeny nejprve podle shluků a potom sestupně pro jednotlivé případy podle šířky obrysu. Zřejmě čím širší obrysy jsou, tím má shlukování lepší vypovídací schopnost.

V R se při fuzzy shlukové analýze vždy vychází z $(n(n-1)/2)$ složkového vektoru nepodobností dvojic případů, pokud je na vstupu zadán namísto toho zdrojový datový soubor s hodnotami jednotlivých sledovaných proměnných pro jednotlivé případy, nejprve se volá procedura pro výpočet nepodobností. Při takovém postupu nemá uživatel možnost změnit přednastavené parametry procedury pro výpočet nepodobností. Možnost zahájit výpočet rovnou s vektorem nepodobností na vstupu může být výhoda při analýzách, kdy datový soubor klasické struktury ani není dostupný.
Při analýze datového souboru metodou fuzzy shlukování pro dva shluky a s přednastavenými hodnotami parametrů ($r = 2$) byly všechny případy nejbližším ostrým přiřazením zařazeny ke správnému shluku, totiž ke správnému zdroji. Dunnův koeficient vychází přibližně 0,8. Uvedeme část (poměrně samovysvětlujícího) výstupu procedury fuzzy shlukování, krácena je ta část výstupu, ve které jsou

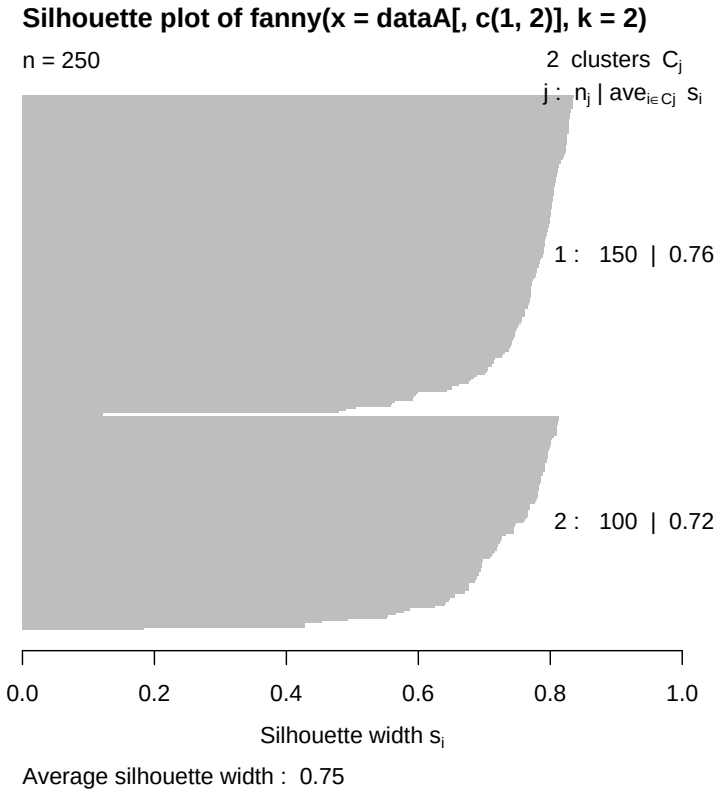
uvedeny příslušnosti případů k jednotlivým shlukům, které zobrazíme pouze pro prvních pět případů.

```
Fuzzy Clustering object of class 'fanny' :
m.ship.expon.               2
objective     186.6242
tolerance        1e-15
iterations          14
converged            1
maxit              500
n                  250



Membership coefficients (in %, rounded):
[,1] [,2]
[1,]   70   30
[2,]   91    9
[3,]   92    8
[4,]   76   24
[5,]   92    8



Fuzzyness coefficients:
dunn_coeff normalized
0.7891172  0.5782343



Closest hard clustering:
[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[149] 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[186] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
[223] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
```

Hodnota účelové funkce (1) se přirozeně bude na stejných datech s rostoucím počtem shluků zlepšovat (jak je také vidět na obrázku (2), který zobrazuje vývoj hodnot účelové funkce (1) v závislosti na počtu shluků k pro k v rozsahu 1 až 5) a dobrou nápovědou o kvalitě nalezeného klasifikačního modelu tak může být spíše graf obrysů shluků (na obrázku (3) zobrazen pro shlukování do dvou shluků), resp. průměrná šířka obrysu, která dosahuje při navrhované klasifikaci do dvou skupin poměrně vysoké hodnoty 0,75. Je dobře patrný prudký pokles hodnoty účelové funkce (1) při rozdělení souboru na dva shluky, při dalším zvyšování počtu shluků se hodnota účelové funkce zlepšuje jen mírně.
Klasifikaci pomocí fuzzy shlukové analýzy lze v tomto případě hodnotit jako dobrou.


Hodnota účelové funkce (1)
mapply(function(x) fanny(dataA[, c(1, 2)], x)$objective, 1:5, SIMPLIFY = TRUE)[1,]
100
200
300
400
500
1
2
3
4
5
k

obr. (2)


Silhouette plot of fanny(x = dataA[, c(1, 2)], k = 2)
n = 250
2 clusters Cj
j : nj | avei∈Cj si
1 : 150 | 0.76
2 : 100 | 0.72
0.0
0.2
0.4
0.6
0.8
1.0
Silhouette width si
Average silhouette width : 0.75

obr. (3)

Metoda shlukování kolem medoidů a CLARA

Metoda CLARA je určena pro zpracování rozsáhlejších datových souborů a je založena na metodě shlukování kolem medoidů (PAM, Partitioning Around Medoids). Krátce podle [1] popíšeme nejprve metodu PAM.

Z množiny n objektů datového souboru se hledá podmnožina k reprezentativních případů (nazvaných medoidy), tedy podle indexů $\{m_1, \ldots, m_k\} \subset \{1, \ldots, n\}$ tak, že je dosaženo pro dané k co nejmenší hodnoty účelové funkce

$$\sum_{i=1}^{n} \min_{t=1,\ldots,k} d(i, m_t) \tag{2}$$

vyjadřující součet nepodobností případu a jeho nejbližšího medoidu přes všechny případy v datovém souboru. Po určení medoidů je přirozeně každý případ přiřazen k nejbližšímu medoidu a každý medoid tak zastupuje jeden shluk.
Vlastní výpočet probíhá ve dvou krocích:

Algoritmus PAM

krok 1: (Vytvoření množiny medoidů)
- polož m_1 je objekt takový, že $\sum_{i=1}^{n} d(i, m_1)$ je minimální
- m_2 je objekt takový, že dojde k největšímu poklesu hodnoty účelové funkce (2)
- $\vdots$
- m_k je objekt takový, že dojde k největšímu poklesu hodnoty účelové funkce (2)

krok 2: (Záměna.)
- Opakuj až do dosažení konvergence: uvažuj všechny dvojice případů (i, j) takové, že i je medoidem a j medoidem není, a proveď záměnu i a j, která nejvíce sníží hodnotu účelové funkce (2), pokud je to ještě možné.

Uvádí se, že metoda PAM by ve srovnání s metodou k-průměrů měla být více robustní z toho pohledu, že medoidy jako zástupci shluků nejsou tolik citlivé k výskytu extrémních hodnot, které vstupují do výpočtu průměru. Střed shluku určený jako (nejspíše fiktivní) objekt s průměrnými hodnotami jednotlivých sledovaných proměnných přes objekty zařazené v daném shluku může být poměrně daleko od skutečných objektů. Navíc je výhodou metody PAM, že není potřeba na úvod zadávat nebo náhodně vybírat množinu reprezentativních případů, jsou totiž v procesu výpočtu PAM poměrně inteligentně nalezeny. Ve srovnání například s fuzzy shlukovou analýzou je výhodou, že metoda PAM, resp. CLARA rovnou poskytuje prostřednictvím medoidů alespoň základní charakteristiku případů zařazených do jednotlivých shluků.

Pro analýzu větších datových souborů (pojem velký datový soubor s růstem výkonnosti dostupných výpočetních prostředků v čase postupně označuje stále větší soubory) je použitelný algoritmus CLARA:

Algoritmus CLARA

- Na vstupu je n případů
- Opakuj x krát:
  - vyber podmnožinu případů o y objektech
  - na danou podmnožinu aplikuj algoritmus PAM a získej množinu medoidů $\{m_1, \ldots, m_k\}$
  - vypočti hodnotu účelové funkce $\sum_{i=1}^{n} \min_{t=1,\ldots,k} d(i, m_t)$
  - uchovej množinu medoidů $\{m_1, \ldots, m_k\}$, pokud je s ní dosaženo zatím nejlepší hodnoty účelové funkce.
- Přiřaď všech n případů k jejich nejbližšímu medoidu z množiny $\{m_1, \ldots, m_k\}$

Při analýze datového souboru metodou CLARA jsou podobně jako v předchozím případě při $k = 2$ objekty zařazovány do správných shluků. Hodnota účelové funkce (2) vydělená počtem případů vychází při práci s eukleidovskou vzdáleností přibližně 1,2. Byly ponechány přednastavené hodnoty parametrů algoritmu, totiž počet výběrů je 5, velikost výběru je $40 + 2 * k$ a vzdálenost je eukleidovská. Ve výstupu jsou mimo jiné uvedeny hodnoty sledovaných proměnných pro oba medoidy a také indexy případů z nejlepšího výběru (tento výběr dosahuje průměrné šířky obrysu 0,79), jehož medoidy jsou nakonec pro shlukování celého souboru určující.

```
Call:        clara(x = dataA[, c(1, 2)], k = 2)
Medoids:
[1,] 10.072661  4.9170248
[2,]  5.002891 -0.1078836
Objective function:            1.241150
Clustering vector:              int [1:250] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 ...
Cluster sizes:                      150 100
Best sample:
[1]   3  23  24  26  45  47  49  55  56  80  83  90  91 101 110 125 130 132 138
[20] 152 154 159 161 167 169 181 185 187 190 195 199 201 206 210 211 212 217 218
[39] 221 233 239 244 245 248
```

Hodnota účelové funkce pro jednotlivé počty shluků se vyvíjí podobně jako při fuzzy shlukové analýze. Výrazně se zlepší po rozdělení souboru do dvou shluků a při dalším zvyšování počtu shluků se zlepšuje jen mírně a tak podle tohoto kritéria je nejspíše rozumné klasifikovat do dvou skupin. Podle očekávání také při zvýšení počtu shluků dochází ke zhoršení profilů shluků v grafu obrysů. Například nejlepší výběr pro $k = 3$ dosahuje průměrné šířky obrysu pouze 0.56. Při klasifikaci do tří skupin je druhý shluk z předchozí analýzy s $k = 2$ ponechán beze změny a první (větší) je rozdělen na dvě skupiny zastoupené 97 a 53 případy.
Použití metody CLARA zřejmě podobně jako při fuzzy shlukové analýze umožňuje nalézt dobrou klasifikaci.

DODATEK: Představení systému R

Systém R je výkonným a flexibilním softwarovým nástrojem a prostředím pro zpracování dat a jejich analýzu, výpočty a tvorbu grafických výstupů. Základem je interpretovaný programovací jayzk s podporou větvení, iterací a modulárního programování pomocí funkcí, jehož návrh vychází z návrhů jazyka S Chamberse a Wilkse a jazyka Scheme a který dává uživateli možnost efektivně definovat funkce pro řešení specifických potřeb. Pro účely zvýšení efektivity výpočtů je navíc možné z prostředí R přistupovat k procedurám vytvořeným v jazycích C, C++ nebo Fortran. Systém R dále obsahuje běhové prostředí a nástroj pro ladění programů a umožňuje spouštět skripty uložené v souborech. Předdefinované funkce pokrývají mnoho statistických postupů například pro lineární modely, zobecněné lineární modely, nelineární regresi, analýzu časových řad, parametrické a neparametrické testy a shlukovou analýzu a k dispozici je rovněž řada doplňkových balíčků zaměřených na některé oblasti analýzy dat. Pro prostředí R existuje podpora importu a exportu datových souborů ve formátech rozšířených statistických a databázových programů. Za pozornost stojí, že jde o software distribuovaný za podmínek licence GNU GPL, což může představovat výraznou výhodu proti běžně dostupným komerčním softwarovým nástrojům pro analýzu dat a statistické výpočty, zejména vzhledem k možnostem modifikace programu a jeho další distribuce a dostupnosti zdrojového kódu. Open source software umožňuje uživateli díky zpřístupnění zdrojového kódu úplnou kontrolu nad postupy použitými při výpočtech (lze-li například zdrojový kód v programovacím jazyce C považovat za dostatečně dobře srozumitelný). Dobrý přehled o použitých algoritmech a detailech jejich implementace je často obtížné získat při využití mnohých komerčních softwarových nástrojů, které, někdy z pochopitelných důvodů, nebývají vždy dostatečně důkladně popsány v dokumentaci dostupné uživatelům.

Literatura:

[1] ROUSSEEUW, P. - STRUYF, A. - HUBERT, M.: Clustering in an Object-Oriented Environment. Journal of Statistical Software, Volume 1, 1996, Issue 4.

[2] R DEVELOPMENT CORE TEAM (2006). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Klíčová slova: shluková analýza, R, silhouette plot.