STATISTICKÉ NÁSTROJE A JEJICH VYUŽITÍ PŘI SEGMENTACI TRHU

STATISTICAL TOOLS AND THEIR UTILIZATION DURING THE PROCESS OF MARKETING SEGMENTATION

Anna Čermáková, Michael Rost

Abstrakt

Cílem příspěvku bylo ukázat, jaké možnosti skýtá využití kofenetického koeficientu korelace při testování shody dat a shlukovacím procesem. Při výpočtech byla použita aglomerativní shlukovací metoda, dvě metriky (Euklidovská a Manhattan) a šest shlukovacích algoritmů. Statistické metody byly použity pro analýzu segmentu dat z marketingového výzkumu. Ukázalo se, že ač metoda nejvzdálenějšího souseda při Euklidovské metrice vykazuje nejvyšší shodu, *CRCC*=0.720, je i takto vysoká shoda neprůkazná. S tímto závěrem koresponduje i výsledek hledání optimálního počtu shluků dle Mojena (1977), který signalizuje jediný možný shluk.

Abstract

The aim of this contribution is to show what possibility are hidden in utilization of the cophenetic coefficient of correlation during the test of the consistency of the data with clustering algorithm. During the computation we used hierarchical agglomerative clustering method with six agglomerative rules and two metrics (Euclidean and Manhattan – city block). This statistical method was used for the analysis of data from marketing survey. It was shown, that although the complete linkage method based on Euclidean metric prove the best consistency, the CRCC =0,720, this consistency is not significant. With this conclusion correspond the result from searching for optimal number of clusters proposed by Mojena (1977). This rules show only one possible cluster .

Klíčová slova:

Shluková analýza, kofenetický koeficient korelace, optimální počet shluků

Key words:

Cluster analysis; cophenetic coefficient of correlation; optimal number of clusters

Úvod

Shluková analýza je zpravidla prováděna na množině objektů, které jsou popsány vektory hodnot statistických znaků. Prostřednictvím této techniky se snažíme zjistit, zda množinu objektů lze rozložit na disjunktní podmnožiny, vnitřně homogenní, avšak navzájem heterogenní. Kvalitní rozklad objektů - např. zákazníků - může napomoci marketingovým manažérům při tvorbě lepších marketingových rozhodnutí a tím vytvoření lepší pozice firmy v konkurenčním prostředí.

Metody a materiál

Jednou z nejčastěji používaných technik shlukové analýzy je aglomerativní hierarchické shlukování. Spočívá v tom, že každý objekt nejprve považujeme za samostatný shluk a poté

objekty či shluky postupně spojujeme na základě propočítané vzdálenosti mezi nimi. Ve finálním stupni shlukování pak všechny objekty tvoří jeden shluk. Shlukujeme vždy ty objekty, které mají v matici vzdáleností nejmenší vzdálenost. Při shlukové analýze musíme řešit tři základní problémy:

1) jakou použít metriku,
2) jak spočítat podobnost nově vzniklého shluku s ostatními objekty či shluky,
3) jaký je „ideální“ počet shluků.

Zabývejme se blíže druhým problémem. Symbolem D označme trojúhelníkovou matici vzdáleností. Maticí vzdáleností rozumíme

a) buď matici vzdáleností mezi objekty - její prvky spočítáme např. prostřednictvím Euklidovské metriky

$$d_{X_i,X_j} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} ,$$

nebo
Manhattan metriky

$$d_{X_i,X_j} = \sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right| ,$$

b) nebo matici vzdáleností mezi shluky. V procesu shlukování se vždy do shluku *t* spojují nejpodobnější shluky (označme je *q, p*), tj. shluky s nejmenší vzdáleností. Spojením shluků vzniká nová situace. Počet shluků se sníží, vzdálenost mezi nimi se musí přepočítat a opět se musí hledat nejvhodnější shluky ke spojení.

Označme: d_{ij} - vzdálenost mezi shluky *i* a *j* (shlukem může být i objekt),

n_i - počet objektů v *i*-tém shluku.

Existuje řada algoritmů pro přepočet prvků nové matice D.

Mezi nejznámější patří:

a) metoda nejbližšího souseda, kde

$$d_{t,r} = \min(d_{p,r}; d_{q,r})$$

b) metoda nejvzdálenějšího souseda, kde

$$d_{t,r} = \max(d_{p,r}; d_{q,r})$$

c) metoda průměrné vazby, kde

$$d_{t,r} = \frac{n_p \cdot d_{p,r} + n_q \cdot d_{q,r}}{n_p + n_q}$$

d) centroidní metoda, kde

$$d_{t,r} = \frac{n_p}{n_p + n_q} \cdot d_{p,r} + \frac{n_q}{n_p + n_q} d_{q,r} - \frac{n_p \cdot n_q}{(n_p + n_q)^2} d_{p,q}$$

e) mediánová metoda, kde

$$d_{t,r} = \frac{d_{p,r} + d_{q,r}}{2} - \frac{d_{p,q}}{4}$$

f) Wardova metoda, kde

$$d_{t,r} = \frac{(n_r + n_p) d_{r,p} + (n_r + n_q) d_{r,q} - n_r \cdot d_{p,q}}{n_t + n_r}$$

Poznámka: v matici D, jejíž prvky jsou na základě předchozí matice D přepočítávány, je řádek a sloupec shluku *q* nově označen jako *t*, a řádek i sloupec shluku *p* jsou vypuštěny.

Vzhledem k počtu možných algoritmů (v kombinaci s různými metrikami) vzniká oprávněná otázka, který z algoritmů vede ke shlukování, jež nejlépe charakterizuje data.

Ačkoli vlastnosti některých shlukovacích algoritmů jsou známy (např. při shlukování prostřednictvím metody nejbližšího souseda se v jednom shluku mohou ocitnout i poměrně vzdálené objekty, metoda nejvzdálenějšího souseda naopak vede k poměrně kompaktním shlukům apod.), je vhodné stupeň shody mezi vlastnostmi objektů a výsledným shlukovacím procesem vyjádřit exaktním ukazatelem. Tímto ukazatelem může být kofenetický koeficient korelace *CPCC* (*Cophenetic Correlation Coefficient*). Jedná se o koeficient korelace mezi prvky primární matice vzdáleností mezi objekty D a mezi prvky kofenetické matice C (*Cophenetix matrix*). Kofenetickou maticí rozumíme trojúhelníkovou matici, jejíž prvky tvoří vzdálenosti mezi shlukovanými objekty v okamžiku, kdy byly poprvé zařazeny do shluku.

Hodnotu kofenetického koeficientu korelace spočítáme podle vztahu:

$$CPCC = \frac{\mathrm{cov}_{d,c}}{s_d \cdot s_c} ,$$

kde *d* jsou prvky primární matice D a *c* jsou prvky kofenetické matice C.

Obecně platí, že čím vyšší je kofenetický koeficient korelace, tím nižší je ztráta informací, vznikající v procesu slučování objektů do shluků.

Rozdělení kofenetického koeficientu při testu hypotéz

H_0: existuje pouze jeden shluk

H_A: existuje systém (množina) kompaktních shluků

studovali F. J. Rohlf a D. L. Fisher (1968). Ukázali, že k zamítnutí H_0 je třeba vysoká hodnota *CPCC*, $CPCC \geq 0,8$. F. J. Rohlf ve své pozdější práci (1970) dokonce doporučuje hodnotu $CPCC > 0,9$.

L. J. Hubert (1974) navrhl pro tento účel použít tzv. Goodman- Kruskalovu statistiku γ.

V případě, že existuje dobrá shoda mezi daty a shlukovacím procesem, je třeba řešit problém uvedený pod bodem 3, tj.zabývat se určením "ideálního počtu shluků. Často lze z dendogramu intuitivně počet shluků odhadnout. Objektivnější postup navrhl R. Mojena (1977). Je založen na relativních velikostech různých shlukovacích úrovní.

Označíme-li shlukovací úrovně $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$, $\alpha_0 < \alpha_1 < \ldots < \alpha_{n-1}$, pak optimální shlukovací úrovní α_{j+1} je první úroveň, pro kterou je

$$\alpha_{j+1} > \bar{\alpha} + ks_\alpha ,$$

$\bar{\alpha}$ je průměr shlukovacích úrovní, s_α je jejich nevychýlená směrodatná odchylka a *k* je číslo z intervalu $\langle 2,75 \cdot 3,50 \rangle$. Pokud takové α neexistuje, soubor objektů tvoří jediný shluk. Problematice určování počtu shluků v hierarchickém shlukování se též věnoval G. W. Milligan (1985).

Cílem příspěvku je ověřit, jak lze kofenetický koeficient využít při výběru vhodného shlukovacího algoritmu a optimálního počtu shluků při zpracování dat marketingového výzkumu.

Data vznikla při dotazníkovém šetření směrujícím k diagnostice nákupních zvyklostí českého zákazníka při nákupu potravin. Z rozsáhlého šetření byly vybrány pouze některé ukazatele, které byly členěny do tříd - viz tab. 1.

Tab. 1: Popis objektu

Třída	Ukazatel
Identifikace zákazníka	Frekvence nákupu Pohlaví Věková kategorie Vzdělání Příjmová kategorie
Umístění prodejny	Blízkost bydliště Blízkost zaměstnání (sídla firmy)
Dostupnost prodejny	Blízkost zastávky MHD Vlastní parkoviště
Sortiment	Výběr téhož výrobku od různých výrobců Kvalita z hlediska chuti, vzhledu, čerstvosti Prodej zákazníkem oblíbeného výrobku
Personál	Ochota a příjemné vystupování Informovanost o novinkách v prodejně
Prodejní doba	Bez poledni přestávky Prodej sedm dní v týdnu
Doplňkové služby	Zákaznické karty Bezhotovostní platba
Ostatní	Rychlost nákupu

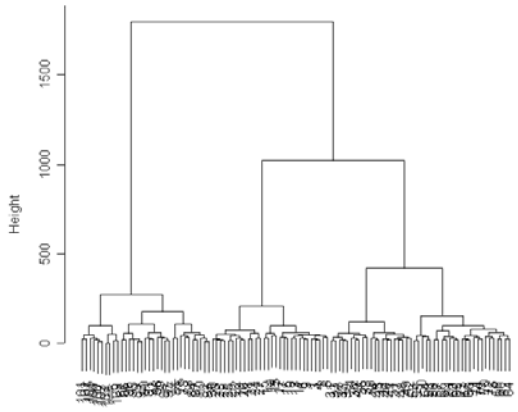
Byly analyzovány informace od 109 zákazníků a použity dvě metriky. Shlukování probíhalo podle šesti shlukovacích algoritmů. Metoda nejbližšího souseda vykazovala velmi nízké hodnoty *CPCC*, není proto mezi výsledky uvedena.

Výsledky

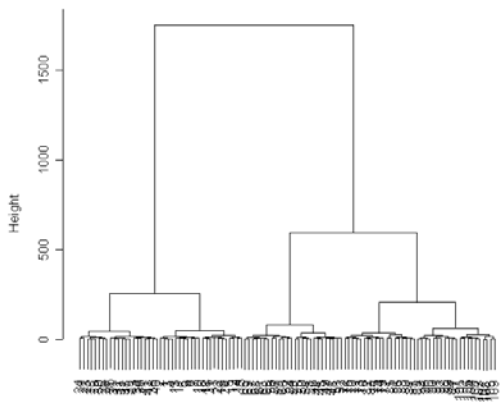
Analýza dat vedla k následujícím výsledkům:

1. *Wardova metoda*

Manhattan metrika,
Kofenetický korel. koeficient: 0.677
Dendrogram:

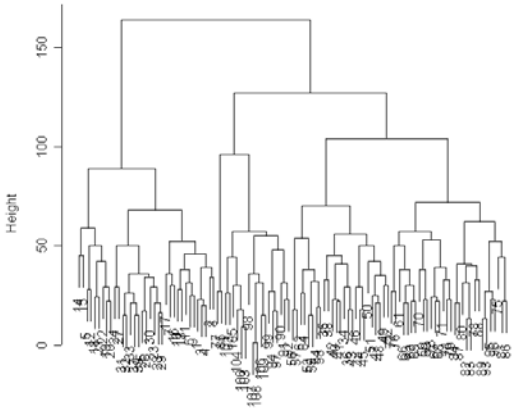


Euklidovská metrika
Kofenetický korel. koeficient: 0.709
Dendrogram:

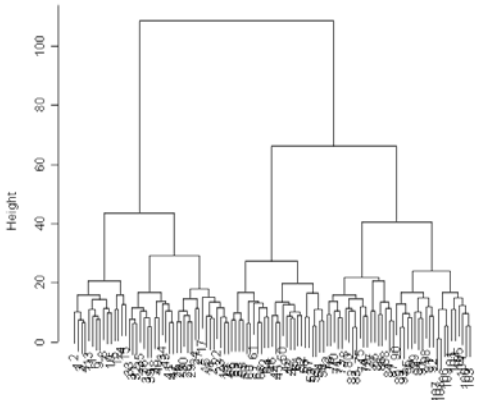


2. *Metoda nejvzdálenějšího souseda*

Manhattan metrika
Kofenetický korel.koeficient: 0.665
Dendrogram:

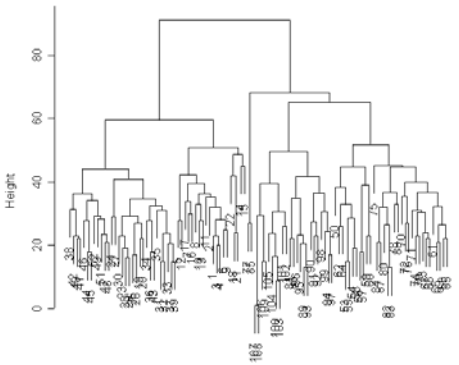


Euklidovská metrika
Kofenetický korel.koeficient:0.720
Dendrogram:

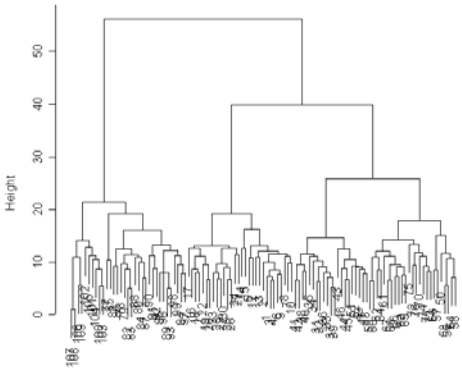


3. *Metoda průměrné vazby*

Manhattan metrika
Kofenetický korel.koeficient: 0.714
Dendrogram:

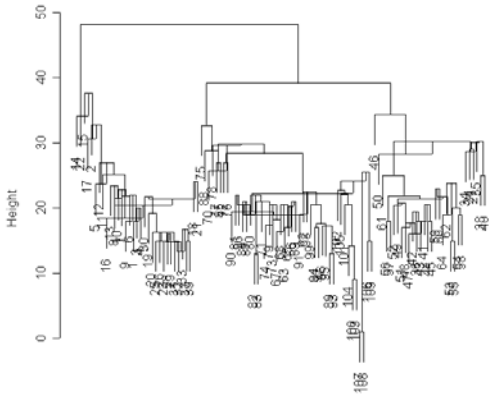


Euklidovská metrika
Kofenetický korel.koeficient: 0.702
Dendrogram:

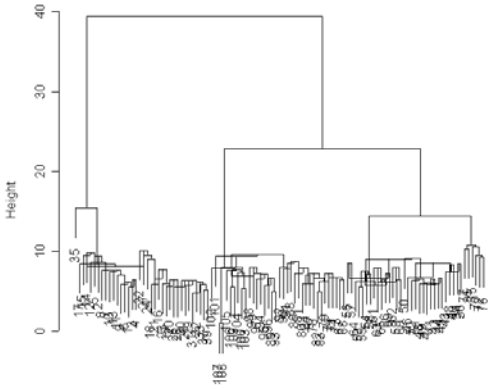


4. *Mediánová metoda*

Manhattan metrika
Kofenetický korel. koeficient: 0.612
Dendrogram:

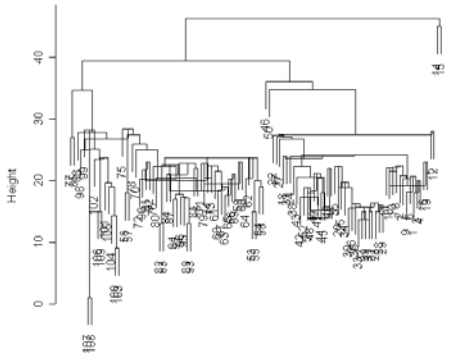


Euklidovská metrika
Kofenetický korel. koeficient: 0.682
Dendrogram:

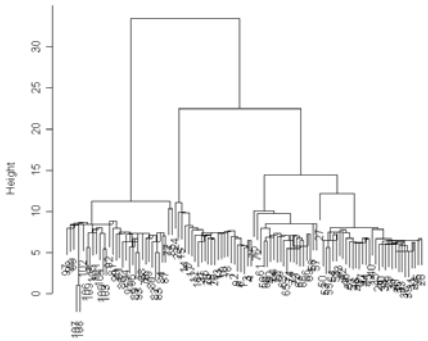


5. *Centroidní metoda*

Manhattan metrika,
Kofenetický korel. koeficient: 0.668
Dendrogram:



Euklidovská metrika
Kofenetický korel. koeficient: 0.694
Dendrogram:



Nejvyšší hodnotu kofenetického koeficientu korelace vykazuje metoda nevzdálenějšího souseda (*CPCC*=0,720). Vzhledem k tomu, že ani takto vysoká hodnota nepřesahuje hodnotu 0,8, nelze segmentaci zákazníků odvozovat na základě příslušného dendogramu. Opticky je tento dendogram velmi ilustrativní a pokud bychom neznali hodnotu *CPCC*, jistě bychom jej použili a snažili bychom se nalézt optimální počet shluků. Na základě dendogramu bychom mohli polemizovat,zda jsou optimální dva, tři či čtyři shluky. Použijeme-li opět statistické metody, konkrétně metodu navrženou Mojenou, získáme tyto výsledky:

$$\alpha_0=0,\ \alpha_1=3,\ldots..,\alpha_{107}=17.378,\ \alpha_{108}=21.587,$$

$$\overline{\alpha}=8.524,\ s_\alpha=3.5,\ \alpha_{j+1}=8{,}524+2.75\cdot 3{,}5=18.15$$

Optimální shlukovací úrovní je tedy α_{108}=21.587, neboť to je první úroveň, přesahující hodnotu 18.15.Dospěli jsme proto k závěru, že zákazníci, diagnostikovaní prostřednictvím 20-ti ukazateli, tvoří jediný shluk.

Výzkumník by si, dospěje-li k takovému závěru, měl mimo jiné položit otázku, zda ukazatele byly vybrány správně.

LITERATURA

Hubert, F. J.(1974): Approximate evaluation/techniques for the single-link and complete-link hierarchical clustering procedures. *Journal of the American Statistical Association,* **69**, pp. 698-704

Milligan, G. W. - Cooper, M. C.(1985): An examination of procedures for determining the number of clusters in a data set. *Psychometrica*, **50**, pp. 159-179

Mojena, R. (1977): Hierarchical grouping methods and stopping rules an evaluation. *Computer Journal*, **20**, pp. 359-364

Rohlf, F. J. (1970): Adaptive hierarchical clustering schemes. *Systematic Zoology*, **19**, pp. 58-82

Rohlf, F. J. - Fisher, D. L.(1968): Test for hierarchical structure in random data sets. *Systematic Zoology*, **17**, pp. 407-412

Adresa autorů

Prof. RNDr. Anna Čermáková, CSc.
katedra aplikované matematiky a informatiky
ZF JU České Budějovice
Studentská 13
370 05
e-meil: annacer@zf.jcu.cz

ing. Michael Rost
katedra aplikované matematiky a informatiky
ZF JU České Budějovice
Studentská 13
370 05
e-meil: rost@zf.jcu.cz