



GRAFICKÁ INTERPRETÁCIA ZHLUKOVEJ ANALÝZY A VÝBER POČTU VÝZNAMNÝCH ZHLUKOV

GRAPHIC INTERPRETATION CLUSTER ANALYSIS AND SELECT THE NUMBER OF IMPORTANT CLUSTER

Jana Halčinová - Peter Trebuňa

Abstract

The article deals with graphical interpretations of cluster analysis hierarchical method, which is tree diagram, so called dendrograph, in vertical or horizontal form. From clusters represented in dendrograph need to be selected optimal amount of clusters. Significant clusters is possible to define with heuristic access or on the basis of quality indicators of clustering.

Key words

Cluster analysis, dendrograph, cluster.

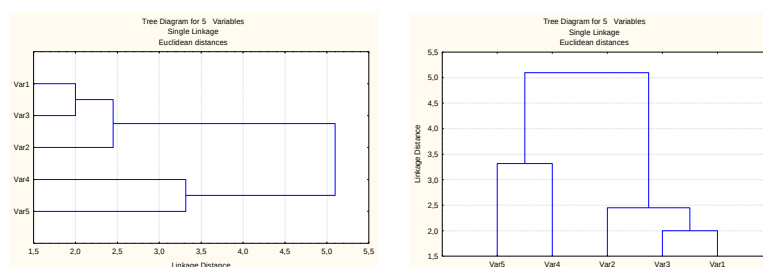
Úvod

Zhluková analýza má široké uplatnenie, predovšetkým však pri zhlukovaní objektov do zhlukov, ktoré majú prirodzenú tendenciu sa zlučovať. Zhlukovou analýzou je možné dosiahnuť usporiadanie objektov do obsahovo podobných zhlukov. Podstatou zhlukovej analýzy je tvorba zhlukov a to na základe podobnosti resp. nepodobnosti zhlukov, pričom objekty v rámci zhluku sú si podobné čo najviac a zhluky navzájom podobné čo najmenej.

Grafická interpretácia zhlukovej analýzy

Výstupom hierarchických metód zhlukovej analýzy môže byť dvojrozmerný diagram známy ako dendrogram, ktorý graficky znázorňuje zhluky na rôznych zhlukovacích úrovniach. Dendrogram je stromový diagram, ktorý sa používa k ilustrácii spojenia vytvoreného v každom kroku analýzy.

Analýzou zhlukov je možné hodnotiť jednak podobnosť objektov – analyzovanou pomocou dendrogramu objektov, alebo podobnosť znakov – analyzovanou pomocou dendrogramu znakov. Dendrogram podobnosti objektov je štandardný výstup hierarchických zhlukovacích metód, z ktorého je zjavná štruktúra objektov v zhlukoch. Dendrogram podobnosti znakov odhaľuje m -tice (najčastejšie dvojice či trojice) znakov, ktoré sú si veľmi podobné a sú vzájomne nahraditeľné. Na obr.1 je znázornený príklad horizontálneho a vertikálneho dendrogramu.



Obr.1 Dendrogram



Výber počtu významných zhlukov

Dendrogram graficky znázorňuje veľa rôznych skupín objektov, ktoré môžu byť prezentované ako zhluky a preto je potrebné stanoviť optimálny počet zhlukov. Existujú dva hlavné prístupy na stanovenie počtu významných zhlukov:

1. heuristický prístup,
2. ukazovatele kvality (efektivity) zhukovania.

Heuristický prístup je najvyužívanejší prístup pre výber významných zhlukov a predstavuje určenie počtu zhlukov na základe subjektívneho názoru riešiteľa. Základným ukazovateľom kvality zhukovania je porovnanie vnútrozhlukového a medzizhlukového rozptylu a to na základe matice vnútrozhlukovej variability:

$$W = \sum_{h=1}^q \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_{C_h}) (x_{hi} - \bar{x}_{C_h})$$

a matice medzizhlukovej variability:

$$B = \sum_{h=1}^q n_h (\bar{x}_{C_h} - \bar{x}) (\bar{x}_{C_h} - \bar{x}),$$

kde $\bar{x}$ je celkový vektor priemerov hodnôt znaku pre celý súbor.

Rozdelenie zhlukov bude optimálne, ak:

- determinant matice vnútrozhlukovej variability bude minimálny,
- stopa matice medzizhlukovej variability bude maximálna.

Kvalitu zhukovania je možné určiť aj pomocou charakteristík uvedených v Tab. 1, ktoré sú taktiež založené „na rozklade sumy štvorcov odchýlok jednotlivých hodnôt od priemeru na medziskupinovú a vnútroskupinovú sumu štvorcov odchýlok počítanú na každom kroku zhukovania“.

Tab. 1 Charakteristiky určujúce počet významných zhlukov

Charakteristiky	Označenie	Popis	Optimálne hodnoty
<i>Štandardná odchýlka</i>	RMSSTD	vyjadruje homogenitu nového zhuku	hodnoty by mali byť nízke
<i>Koeficient determinácie</i>	RSQ	vyjadruje heterogenitu zhlukov	hodnoty by mali byť vysoké
<i>Semiparciálny koeficient determinácie</i>	SPRSQ	vyjadruje homogenitu spojených zhlukov	hodnoty by mali byť nízke
<i>Vzdialenosť zhlukov</i>	CD	vyjadruje homogenitu spojených zhlukov	hodnoty by mali byť nízke

Súhrn

Zhluk predstavuje skupinu objektov, ktorých vzdialenosť tzn. nepodobnosť je menšia ako vzdialenosť objektov do zhuku nepatriacich. Zhluky hierarchických metód zhukovej analýzy sú prezentované vo forme dendrogramov, ktoré graficky znázorňujú rôzne skupiny zhlukov v rôznych zhukovacích úrovniach. Najčastejšie sú významné zhluky volené na základe subjektívneho názoru, tzn. pomocou heuristického prístupu. Iným spôsobom výberu významných zhlukov je stanovenie charakteristík určujúcich počet významných zhlukov ako sú štandardná odchýlka, koeficient determinácie, semiparciálny koeficient determinácie a vzdialenosť zhlukov.



Kľúčové slová

Zhluková analýza, dendrogram, zhluk.

Použitá literatúra

- [1] MELOUN, M. – MILITKÝ, J. – HILL, M.: *Počítačová analýza vícerozměrných dat v příkladech*. Praha: Academia. 2005. 449 s. ISBN 80-200-1335-0,
- [2] MELOUN, M. – MILITKÝ, J.: *Statistická analýza experimentálních dat*. Praha: Academia. 2004. 953 s. ISBN 80-200-1254-0,
- [3] ŘEZANKOVÁ, H. – HÚSEK, D. – SNÁŠEL, V.: *Shluková analýza dat*. Praha: Professional Publishing. 2009. 218 s. ISBN 978-80-86946-81-8,
- [4] STANKOVIČOVÁ, I. – VOJTKOVÁ, M.: *Viacrozmerné štatistické metódy s aplikáciami*. Bratislava: IURA Edition, 2007, 261 s. ISBN 978-80-8078-152-1.

Kontaktná adresa

Ing. Jana Halčinová
doc. Ing. Peter Trebuňa, PhD.
Technická univerzita v Košiciach
Strojnícka fakulta
Katedra priemyselného inžinierstva a manažmentu
Němcovej 32, 040 01 Košice
e-mail: jana.halcinova@tuke.sk,
peter.trebuna@tuke.sk