# PRINCÍPY ZHLUKOVEJ ANALÝZY

# PRINCIPLES OF CLUSTER ANALYSIS

Peter TREBUŇA - Jana HALČINOVÁ

**Abstract**

The main point of this article is a brief description of principle of cluster analysis. The cluster analysis is a set of methods, which help us to search in empirical datas cluster of similary objects. To study similarity of objects are used the degrees of similarity, and more often degrees of unsimiliraty. The graphical output of objects cluster is a dendrograph.

**Key words**

cluster analysis, cluster, degree of similarity, methods of cluster analysis, dendrograph

## Úvod

Zhluková analýza patrí medzi metódy, ktoré sa zaoberajú podobnosťou viacrozmerných objektov a klasifikáciou objektov do zhlukov. Zhlukovú analýzu môžeme vo všeobecnosti definovať ako všeobecný logický postup, formulovaný ako procedúra, pomocou ktorej sa zlučujú objekty do skupín - zhlukov, a to na základe ich podobnosti a rozdielnosti. Zhlukovú analýzu je možné využiť aj pre radikálne zníženie dimenzie úlohy a to tak, že uvažované premenné zastúpi jediná premenná vyjadrujúca príslušnosť k takto definovanému zhluku. Zhluk je skupina objektov, ktorých vzdialenosť (nepodobnosť) je menšia ako vzdialenosť, ktorú majú objekty do zhluku nepatriace.

## Miery podobnosti

Miera podobnosti pre objekty $x_i$ a $x_j$ sa zapisuje ako $S(x_i, x_j)$, skrátene $S_{ij}$ a platí, že $S_{ij} = S_{ji}$. Miery podobnosti v ideálnom prípade nadobúdajú hodnoty z intervalu <0,1>, pričom 0 vyjadruje maximálnu rozdielnosť objektov a hodnota 1 maximálnu totožnosť. Miera nepodobnosti objektov $x_i$ a $x_j$ sa zapisuje ako $D(x_i, x_j)$, skrátene $D_{ij}$ a platí, že:

- $D_{ij} \geq 0$,
- $D_{ii} = 0$,
- $D_{ij} = D_{ji}$.

Podobnosť objektov môže byť meraná rôznymi spôsobmi, ktoré sa dajú obyčajne zaradiť do týchto základných skupín:

- miery asociácie,
- miery vzdialenosti (metriky),
- miery korelácie,

pričom koeficienty asociácie a korelácie predstavujú miery podobnosti objektov a metriky predstavujú miery nepodobnosti objektov.

***Miery asociácie*** je skupina koeficientov určená pre hodnotenie podobnostných vzťahov objektov charakterizovaných dichotomickými (nemetrickými) znakmi. Asociáciu dvojice objektov $x_i$ a $x_j$ charakterizuje asociačná tabuľka (tab.1)

Tab. 1 Asociačná tabuľka

| | | $x_i$ | |
|---|---|---|---|
| | | 1 | 0 |
| $x_j$ | 1 | $a$ | $b$ |
| | 0 | $c$ | $d$ |

V tab.1 sú obsiahnuté všetky možné kombinácie počtu znakov pre objekty $x_i$ a $x_j$, kde:

$a$ je počet znakov, keď majú objekty $x_i$ a $x_j$ hodnotu 1 a ide o tzv. pozitívnu zhodu,

$b$ – počet znakov, keď má objekt $x_j$ hodnotu 1 a objekt $x_i$ hodnotu 0,

$c$ – počet znakov, keď má objekt $x_j$ hodnotu 0 a objekt $x_i$ hodnotu 1,

$d$ – počet znakov, keď majú objekty $x_i$ a $x_j$ hodnotu 0 a ide o tzv. negatívnu zhodu.

Medzi základné miery asociácie patria:

**a)** *Sokalov – Michenerov koeficient asociácie (koeficient jednoduchej zhody)*

$$S_J = \frac{a}{a+b+c}, \tag{1}$$

**b)** *Russelov – Raoov koeficient asociácie*

$$S_H = \frac{a+d-b-c}{a+b+c+d}, \tag{2}$$

**c)** *Jaccardov koeficient*

$$S_{RT} = \frac{a+d}{a+2b+2c+d}, \tag{3}$$

**d)** *Hamannov koeficient asociácie*

$$S_S = \frac{2a}{2a+b+c}, \tag{4}$$

**e)** *Rogerosov a Tanimotov koeficient asociácie*

$$S_{SM} = \frac{a+d}{a+b+c+d}, \tag{5}$$

**f)** *Sörensenov koeficient asociácie*

$$S_{RR} = \frac{d}{a+b+c+d}. \tag{6}$$

***Miery vzdialenosti*** predstavujú najpoužívanejšie miery založené na prezentácii objektov v priestore, ktorého súradnice tvoria jednotlivé premenné a využívajú sa v štatistických programoch. Ak je splnená trojuholníková nerovnosť $D_{iy} + D_{yj} \geq D_{ij}$ a $(i, j, y \in < 1; n >)$, kde $n$ predstavuje počet objektov, potom hovoríme o ***metrike***.

K najznámejším typom miery vzdialenosti, resp. metriky patria:

**a)** *Euklidovská vzdialenosť*

$$D_E(x_i, x_j) = \sqrt{\sum_{l=1}^{m} (x_{il} - x_{jl})^2} = \left\| x_i - x_j \right\|, \tag{7}$$

**b)** *Vážená euklidovská vzdialenosť*

$$D_{EW}(x_i, x_j) = \sqrt{\sum_{l=1}^{m} w_l (x_{il} - x_{jl})^2}, \tag{8}$$

kde $w_l$ sú váhy pre každú $l$-tú premennú,

**c)** *Štvorcová euklidovská vzdialenosť*

$$D_{ES}(x_i, x_j) = \sum_{l=1}^{m} (x_{il} - x_{jl})^2 , \qquad (9)$$

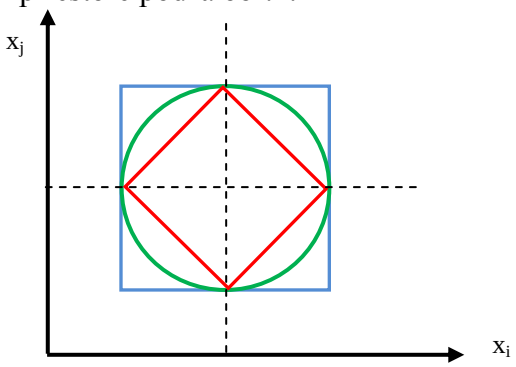
**d)** *Manhattanská (mestských blokov) vzdialenosť*

$$D_B(x_i, x_j) = \sum_{l=1}^{m} \left| x_{il} - x_{jl} \right| = \left| x_i - x_l \right| , \qquad (10)$$

**e)** *Čebyševova vzdialenosť*

$$D_C(x_i, x_j) = \max_l \left( \left| x_{il} - x_{jl} \right| \right), \qquad (11)$$

Ak sú objekty $x_i$ a $x_j$ charakterizované dvoma premennými, môžeme ich vzdialenosť znázorniť v dvojrozmernom priestore podľa obr.1.




Obr. 1 Vzdialenosť objektov v dvojrozmernom priestore

Z obr.1 je zrejmé, že pre danú mieru majú objekty od stredu obrazca až po jeho obvod rovnakú vzdialenosť, sú osovo symetrické. Vnútorný štvorec charakterizuje manhattanskú, kruh euklidovskú a vonkajší štvorec Čebyševovu vzdialenosť.

Základnou ***mierou korelácie*** je *Pearsonov korelačný koeficient*, ktorý sa pre *k*-tú a *l*-tú premennú určí nasledovne:

$$r_{kl} = \frac{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)(x_{il} - \bar{x}_l)}{\sqrt{\sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2 \sum_{i=1}^{n} (x_{il} - \bar{x}_l)^2}} , \qquad (12)$$

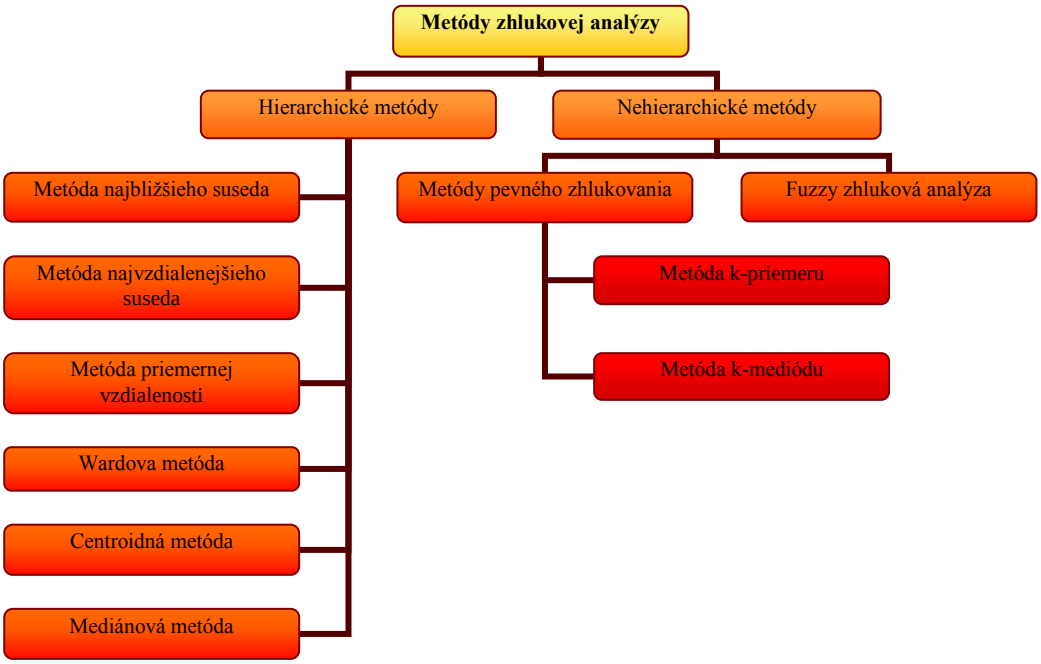
kde:

$\bar{x}_k$ - je aritmetický priemer hodnôt *k*-tej premennej,

$\bar{x}_l$ - je aritmetický priemer hodnôt *l*-tej premennej.

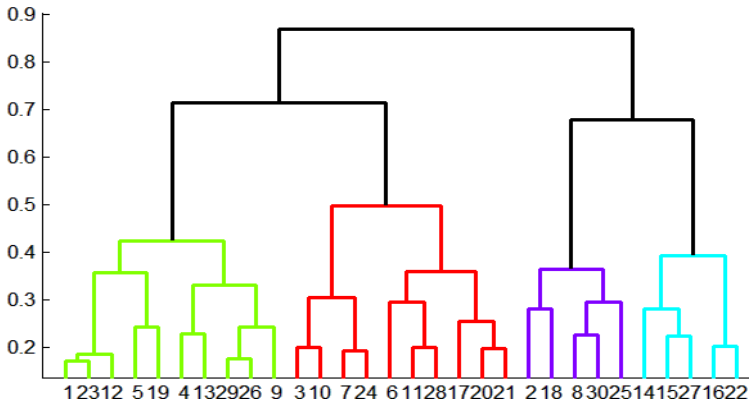
**Metódy zhlukovej analýzy**

Pod pojmom zhluková analýza rozumieme skupinu metód, ktorých cieľom je buď zoskupiť dané objekty do zhlukov (hierarchické metódy) alebo vytvoriť hierarchiu zhlukov objektov (nehierarchické). Klasifikácia metód zhlukovej analýzy je zobrazená na obr.2.




Obr. 2 Klasifikácia metód zhlukovej analýzy

Dendrogram (obr.3) je grafická prezentácia hierarchicky usporiadaných zhlukov formou vývojového stromu.




Obr. 3 Dendrogram

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

1. Heuristický prístup
2. Ukazovatele kvality (efektivity) zhlukovania

Heuristický prístup je využívaný najčastejšie a predstavuje určenie počtu zhlukov na základe subjektívneho názoru riešiteľa.

Základným ukazovateľom kvality zhlukovania je porovnanie vnútrozhlukového a medzizhlukového rozptylu jednotlivých premenných a to na základe matice vnútrozhlukovej variability:

$$W = \sum_{h=1}^{q} \sum_{i=1}^{n_h} (x_{hi} - \bar{x}_{C_h})' (x_{hi} - \bar{x}_{C_h}) \qquad (13)$$

a matice medzizhlukovej variability

$$B = \sum_{h=1}^{q} n_h (\bar{x}_{C_h} - \bar{x})' (\bar{x}_{C_h} - \bar{x}), \qquad (14)$$

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

Rozdelenie zhlukov bude optimálne, ak:

- determinant matice vnútrozhlukovej variability bude minimálny (Wardovo kritérium kvality),
- stopa matice medzizhlukovej variability bude maximálna.

**Súhrn**

Príspevok stručne popisuje princíp zhlukovej analýzy, ktorá predstavuje súbor metód určených pre exploračnú analýzu dát. Metódy majú za cieľ v množine analyzovaných dát nájsť určité skryté štruktúry a prezentovať ich ako skupiny – zhluky podobných objektov. Podobnosť objektov je teda funkciou vlastností objektov. Základom zhlukovej analýzy je vytvoriť zhluky objektov tak, aby objekty v určitom zhluku si boli navzájom podobné čo najviac a čo najmenej podobné objektom v iných zhlukoch. Grafickým znázornením rôznych skupín zhlukov je dendrogram, z ktorého prostredníctvom heuristického prístupu resp. ukazovateľov kvality zhlukovania je možné stanoviť optimálny počet zhlukov.

**Kľúčové slová**

zhluková analýza, zhluk, miery podobnosti, metódy zhlukovej analýzy, dendrogram

*Tento článok bol vytvorený realizáciou projektu "Centrum výskumu riadenia technických, environmentálnych a humánnych rizík pre trvalý rozvoj produkcie a výrobkov v strojárstve" (ITMS: 26220120060), na základe podpory operačného programu Výskum a vývoj financovaného z Európskeho fondu regionálneho rozvoja..*

**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.
[5] TREBUŇA, P., BÉREŠ, M.: Klasifikácia metód zhlukovania a oblasti ich využitia 2010. In: Transfer inovácií. - ISSN 1337-7094. - Č. 16 (2010), s. 31-34

**Kontaktná adresa**

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