

ALGORITMUS METÓDY K-PRIEMEROV

ALGORITHM OF THE METHOD K-MEANS

Peter TREBUŇA - Jana HALČINOVÁ

Abstrakt

Príspevok je zameraný na podstatu a postup najpoužívanejšej nehierarchickej metódy zhlukovej analýzy, metódy k-priemerov. V článku je popísaný algoritmus tejto metódy, ktorý je aplikovaný na dátovú maticu o rozmeroch 5×2 tzn. 5 objektov charakterizovaných 2 znakmi.

Kľúčové slová: zhluková analýza, nehierarchické metódy, metóda k-priemerov

Abstract

The article is focused on the very core and procedure of mostly used non-hierarchical method of cluster analysis, k-means method. This article describes an algorithm of the method, which is applied on data matrix 5×2 ie. 5 objects characterized 2 features.

Key words: cluster analysis, non-hierarchical methods, method k-means

ÚVOD

Metódy zhlukovej analýzy sa vyvinuli z potreby analyzovať informácie obsiahnuté v množine údajov, o ktorých štruktúre vieme len veľmi málo. Štruktúrou sa pritom rozumie rozdelenie objektov do určitého systému skupín, ktoré zohľadňujú podobnosť objektov v skupine a nepodobnosť objektov patriacich do rôznych skupín. Metódy zhlukovej analýzy poskytujú algoritmy, ktoré umožňujú takúto klasifikáciu objektov do skupín – zhlukov a členia sa na hierarchické a nehierarchické metódy, pričom v článku je pozornosť upriamená na jednu z nehierarchických metód zhlukovania.

Podstata metódy k-priemerov

Metóda k-priemerov (k-means) patrí medzi najjednoduchšie a najrozšírenejšie nehierarchické metódy, ktoré nemenia zadaný počet zhlukov.

Postup zhlukovania:

1. Zvolíme si k počiatkových objektov $A_1^{(0)}, A_2^{(0)}, \dots, A_k^{(0)}$, ktoré môžu byť náhodne vybrané z množiny objektov. Obyčajne sa volí prvých k objektov.
2. V prvej iterácii vytvoríme rozklad množiny objektov $S_1 = \{C_1^{(1)}, C_2^{(1)}, \dots, C_{l_1}^{(1)}\}$ na zhluky tak, aby pre každý objekt $A_j^{(1)} \in C_i^{(1)}$ platilo:

$$A_j^{(1)} \in C_i^{(1)} \Leftrightarrow d(A_j^{(1)}, A_i^{(1)}) = \min_{m \in \{1, 2, \dots, k\}} \{d(A_j^{(1)}, A_m^{(0)})\}.$$
3. Pre každý zhluk $C_i^{(1)}$ určíme jeho centroid $A_i^{(1)}$ pre $i=1, 2, \dots, l_1$. Dostaneme tak množinu nových centier, pre ktoré sa vypočíta nový rozklad $S_2 = \{C_1^{(2)}, C_2^{(2)}, \dots, C_{l_2}^{(2)}\}$.
4. Nech v s -tej iterácii je rozklad $S_s = \{C_1^{(s)}, C_2^{(s)}, \dots, C_{l_s}^{(s)}\}$ a zodpovedajúce centrá $A_1^{(s)}, A_2^{(s)}, \dots, A_{l_s}^{(s)}$. Potom ďalší rozklad $S_{s+1} = \{C_1^{(s+1)}, C_2^{(s+1)}, \dots, C_{l_{s+1}}^{(s+1)}\}$ vytvoríme tak, aby pre každý objekt $A_j^{(s+1)}$ platilo

$$A_j^{(s+1)} \in C_i^{(s+1)} \Leftrightarrow d(A_j^{(s+1)}, A_i^{(s)}) = \min_{m \in \{1, 2, \dots, k\}} \{d(A_j^{(s+1)}, A_m^{(0)})\}.$$
5. Po získaní každého ďalšieho rozkladu sa porovnávajú dva za sebou idúce rozklady medzi sebou. Môžu nastať tieto prípady:
 - ak sa aspoň jedným zhlukom tieto rozklady od seba líšia, potom sa pokračuje v ďalšom prerozdelení objektov do zhlukov,
 - ak sa zhluky v oboch rozkladoch rovnajú, potom zhlukovací proces sa končí a rozklad $S_{s+1} = \{C_1^{(s+1)}, C_2^{(s+1)}, \dots, C_{l_{s+1}}^{(s+1)}\}$ tvorí výsledok zhlukovania.

Postup metódy k-priemerov

Vstupom pre použitie metódy k -priemerov je dátová matica (Tab.1). Ako miera vzdialenosti je použitá štvorcová euklidovská vzdialenosť a postupuje sa podľa popísaného algoritmu.

Prvým krokom procesu zhľukovania je rozklad dát na $k=2$ zhľuky.

Jednotlivé zhľuky sú $C_1 = \{A_1, A_2, A_3\}$ a $C_2 = \{A_4, A_5\}$.

Tab. 1 Vstupná dátová matica

Objekt	1. znak	2. znak
A_1	1	1
A_2	1	2
A_3	6	3
A_4	8	2
A_5	8	0

V ďalšom kroku je potrebné určiť centroidy týchto zhľukov:

$$K_1 = \left(\frac{1+1+6}{3}; \frac{1+2+3}{3} \right) = (2,7;2)$$

$$K_2 = \left(\frac{8+8}{2}; \frac{2+0}{2} \right) = (8;1)$$

Nasleduje výpočet vzdialenosti jednotlivých objektov od centroidov:

$$\begin{aligned} d_{es}(K_1, A_1) &= (2,7-1)^2 + (2-1)^2 = 3,89 \\ d_{es}(K_2, A_1) &= (8-1)^2 + (1-1)^2 = 49 \end{aligned} \quad \begin{matrix} > \\ > \end{matrix} C_1$$

Pretože je vzdialenosť $d_{es}(K_1; A_1) < d_{es}(K_2; A_1)$, zostáva tento objekt v zhľuku C_1 . Analogické výpočty sa uskutočnia aj pre zvyšné objekty.

$$\begin{aligned} d_{es}(K_1, A_2) &= (2,7-1)^2 + (2-2)^2 = 2,89 \\ d_{es}(K_2, A_2) &= (8-1)^2 + (1-2)^2 = 50 \\ d_{es}(K_1, A_3) &= (2,7-6)^2 + (2-3)^2 = 11,89 \\ d_{es}(K_2, A_3) &= (8-6)^2 + (1-3)^2 = 8 \\ d_{es}(K_1, A_4) &= (2,7-8)^2 + (2-2)^2 = 28,09 \\ d_{es}(K_2, A_4) &= (8-8)^2 + (1-2)^2 = 1 \\ d_{es}(K_1, A_5) &= (2,7-8)^2 + (2-0)^2 = 32,09 \\ d_{es}(K_2, A_5) &= (8-8)^2 + (1-0)^2 = 1 \end{aligned} \quad \begin{matrix} > \\ > \\ > \\ > \\ > \\ > \end{matrix} \begin{matrix} C_1 \\ C_1 \\ C_1 \\ C_2 \\ C_2 \\ C_2 \end{matrix} \quad \begin{matrix} \\ \\ C_1 - \text{presun do } C_2 \\ \\ \end{matrix}$$

Na základe presunov vznikli tieto nové zhľuky $C_1 = \{A_1, A_2\}$ a $C_2 = \{A_3, A_4, A_5\}$ a preto sa opakuje postup prepočtu vzdialenosti objektov od nových centroidov. Nové centroidy (ťažiská) zhľukov dosahujú hodnoty:

$$K_1 = \left(\frac{1+1}{2}; \frac{1+2}{2} \right) = (1;1,5)$$

$$K_2 = \left(\frac{6+8+8}{3}; \frac{3+2}{3} \right) = (7,3;1,7)$$

Od týchto centroidov sa určí vzdialenosť jednotlivých objektov a zhodnotí sa, či dochádza k presunu jednotlivých objektov do iných zhlukov:

$$\begin{aligned} d_{ES}(K_1, A_1) &= (1-1)^2 + (1,5-1)^2 = 0,25 \\ d_{ES}(K_2, A_1) &= (7,3-1)^2 + (1,7-1)^2 = 40,18 \\ d_{ES}(K_1, A_2) &= (1-1)^2 + (1,5-2)^2 = 0,25 \\ d_{ES}(K_2, A_2) &= (7,3-1)^2 + (1,7-2)^2 = 39,78 \\ d_{ES}(K_1, A_3) &= (1-6)^2 + (1,5-3)^2 = 27,25 \\ d_{ES}(K_2, A_3) &= (7,3-6)^2 + (1,7-3)^2 = 3,38 \\ d_{ES}(K_1, A_4) &= (1-8)^2 + (1,5-2)^2 = 49,25 \\ d_{ES}(K_2, A_4) &= (7,3-8)^2 + (1,7-2)^2 = 0,58 \\ d_{ES}(K_1, A_5) &= (1-8)^2 + (1,5-0)^2 = 51,25 \\ d_{ES}(K_2, A_5) &= (7,3-8)^2 + (1,7-0)^2 = 3,38 \end{aligned}$$

 $\rangle C_1$
 $\rangle C_1$
 $\rangle C_2$
 $\rangle C_2$
 $\rangle C_2$

Nakoľko v tejto etape zhlukovacieho procesu už nedochádza k ďalším presunom objektov do zhlukov, považujú sa zhluky $C_1 = \{A_1, A_2\}$ a $C_2 = \{A_3, A_4, A_5\}$ za výsledné riešenie zhlukovej analýzy metódou k -priemerov.

ZÁVER

Metódy zhlukovej analýzy teda slúžia ku klasifikácii objektov do zhlukov na základe podobnosti resp. nepodobnosti, pričom podmienkou je, aby si objekty v rámci zhliku boli podobné čo najviac a zhluky navzájom podobné čo najmenej. Hierarchické metódy zhlukovej analýzy zaraďujú analyzované objekty do hierarchického systému zhlukov. Pri nehierarchických metódach zhlukovej analýzy ide o zaradenie objektov do vopred stanoveného počtu disjunktných zhlukov.

♦♦ *Príspevok bol pripravený v rámci riešenia grantového projektu VEGA č. 1/0102/11 Metódy a techniky experimentálneho modelovania vnútropodnikových výrobných a nevýrobných procesov.*

ZOZNAM BIBLIOGRAFICKÝCH ODKAZOV

- [1] EVERITT, B.S. – LANDAU, S. – LEESE, M. – STAHL, D.: Cluster Analysis. - Londýn: Wiley, 2011. 348s. ISBN 978-0-470-74991-3.
- [2] HANDL, A.: Multivariate Analysemethoden – Theorie und Praxis multivariater Verfahren unter besonderer Berücksichtigung von S-Plus. Berlin: Springer. 2002.
- [3] HEBÁK, P.: Vícerozměrné statistické metody s aplikacemi. - Praha: SNTL/ALFA, 1987. 456 s. 04-323-87.
- [4] LUKASOVÁ, A. – ŠARMANOVÁ, J.: Metody shlukové analýzy. Praha: SNTL, 1985. 212 s. 04-014-85.
- [5] MARKOVIČ, J., JAMRICHOVÁ, S.: Prosperující podnik v globálním prostředí. - Invent, Žilina 2008
- [6] MARKOVIČ, J., KUČERA, Ľ., LUKÁČ, M.: Metrológia ako nástroj zvýšenia bezpečnosti cestnej premávky. - Rosaline, Bratislava 2009
- [7] TREBUŇA, P. – HALČINOVÁ, J.: Experimental modelling of the cluster analysis processes. - In: Procedia Engineering. Vol. 48. 2012, s. 673–678. ISSN 1877-7058.

ADRESY AUTOROV

Peter TREBUŇA, doc. Ing., PhD., Technická univerzita v Košiciach, Strojnícka fakulta, Katedra priemyselného inžinierstva a manažmentu, Němcovej 32, 042 00 Košice, Slovenská republika, e-mail: peter.trebuna@tuke.sk

Jana HALČINOVÁ, Ing., Technická univerzita v Košiciach, Strojnícka fakulta, Katedra priemyselného inžinierstva a manažmentu, Němcovej 32, 042 00 Košice, Slovenská republika, e-mail: jana.halcinova@tuke.sk

RECENZENT

Vojtech KOLLÁR, prof. Ing., PhD., Vysoká škola ekonómie a manažmentu verejnej správy v Bratislave, Furdekova 16, 851 04 Bratislava 5, Slovenská republika