MĚŘENÍ PODOBNOSTI OBJEKTŮ A SHLUKŮ PŘI SHLUKOVÉ ANALÝZE S KVALITATIVNÍMI PROMĚNNÝMI A PROMĚNNÝMI RŮZNÝCH TYPŮ

Tomáš Löster

Abstrakt

Shluková analýza je vícerozměrná statistická metoda, jejíž cílem je vytvářet množiny objektů, tzv. *shluky*, v rámci kterých by si objekty (vícerozměrná pozorování charakterizovaná řadou vlastností) měly být co nejvíce podobné z hlediska vnitroshlukové struktury a co nejméně podobné z hlediska mezishlukové struktury. Shlukovat lze také proměnné, případně kategorie nominálních proměnných, případně objekty i proměnné současně. Shluková analýza je využívána v řadě vědních oborů, mj. v demografii. Klíčovou úlohu ve shlukové analýze zaujímá stanovení podobnosti objektů, přičemž je potřeba rozlišit, jakými typy proměnných jsou vlastnosti jednotlivých objektů charakterizovány. Mohou to být proměnné kvantitativní, kvalitativní (nominální nebo ordinální), nebo proměnné různých typů (kombinace kvantitativních a kvalitativních proměnných). Cílem tohoto příspěvku je popsat možnosti měření podobnosti objektů a shluků v případě, jsou-li objekty charakterizovány proměnnými různých typů.

Klíčová slova: Shluková analýza, podobnost objektů, podobnost shluků

JEL Code: C3, C38, C40

Úvod

Klíčovou úlohu ve shlukové analýze zaujímá stanovení podobnosti objektů a shluků, přičemž je potřeba rozlišit, jakými typy proměnných jsou vlastnosti jednotlivých objektů charakterizovány. Mohou to být proměnné kvantitativní, kvalitativní (nominální nebo ordinální), nebo proměnné různých typů. Zvláštním případem jsou dichotomické proměnné, které nabývají pouze dvou hodnot. Obvykle jsou to hodnoty 0 a 1 a proměnné se označují jako *binární*. V případě, že jsou objekty charakterizovány pouze kvantitativními proměnnými, v současné literatuře existuje mnoho koeficientů, které vychází především z měr vzdáleností, viz např. [1]. Mezi tyto míry patří například Euklidovská či Manhattanská vzdálenost.

1 Měření podobnosti objektů

V případě, že jsou objekty charakterizovány pomocí *m* proměnných různých typů, pak je při měření podobnosti dvou objektů využíván **Gowerův koeficient podobnosti**, viz [4], který je definován jako

$$A_{\mathrm{GW}} = \frac{\sum_{t=1}^{m} w_{ijt} A_{ijt}}{\sum_{t=1}^{m} w_{ijt}}, \qquad (1)$$

kde w_{ijt} nabývá hodnot 0 (jestliže hodnota *t*-té proměnné u *i*-tého nebo *j*-tého objektu chybí nebo jsou obě tyto hodnoty rovny nule a *t*-tá proměnná je binární), nebo 1 (v ostatních případech).

Míra podobnosti A_{ijt} závisí na typu *t*-té proměnné. V případě, že *t*-tá proměnná je binární nebo nominální, pak

$$A_{ijt} = 1 \text{ pro } x_{it} = x_{jt}, \qquad (2)$$

$$A_{ijt} = 0 \text{ jinak.} \qquad (3)$$

V případě, že *t*-tá proměnná je kvantitativní pak

$$A_{ijt} = 1 - \frac{\left| x_{it} - x_{jt} \right|}{R_t}, \qquad (4)$$

kde R_t je variační rozpětí *t*-té proměnné určené na základě celého souboru.

Dva objekty jsou si nejpodobnější v případě, že shluk z nich vytvořený vykazuje nejmenší variabilitu. K měření variability lze použít rozptyl v kombinaci s entropií. Variabilitu *h*-tého shluku lze stanovit podle vzorce

$$H_h = \sum_{t=1}^{m_1} \frac{1}{2} \ln(s_t^2 + s_{ht}^2) + \sum_{t=1}^{m_2} H_{ht}, \qquad (5)$$

kde m_1 je počet kvantitativních (spojitých) proměnných, m_2 je počet nominálních proměnných, s_t^2 je vyběrový rozptyl *t*-té proměnné a s_{ht}^2 je výběrový rozptyl *t*-té proměnné v *h*-tém shluku, kde míra variability nominální proměnné pro *t*-tou proměnnou v *h*-tém shluku se určí jako

$$H_{ht} = -\sum_{u=1}^{K_t} \left(\frac{n_{htu}}{n_h} \ln \frac{n_{htu}}{n_h} \right), \qquad (6)$$

kde K_t je počet kategorií t-té proměnné, je n_{htu} je počet objektů u-té kategorie, t-té proměnné v h-tém shluku a n_h je počet objektů v h-tém shluku.

Tento postup je využit ve dvoukrokové shlukové analýze v systému SPSS. Ta je navržena pro shlukování velkého počtu objektů a je založena na algoritmu BIRCH, v němž jsou objekty uspořádány do podshluků, které jsou charakterizovány pomocí shlukovacích vlastností, viz [5].

V případě, že jsou objekty charakterizovány pomocí kombinace kvantitativních a nominálních proměnných, je možné navrhnout míru variability s využitím rozptylu a hodnot Giniho koeficientu, tedy podle vzorce

$$G_h = \sum_{t=1}^{m_1} \frac{1}{2} \ln(s_t^2 + s_{ht}^2) + \sum_{t=1}^{m_2} G_{ht} \,. \qquad (7)$$

Dále pro případ, kdy jsou objekty charakterizovány kvantitativními a ordinálními proměnnými, je navíc možné navrhnout míru variability s využitím rozptylu a koeficientu dorvar. Vypočítá se podle vzorce

$$DK_{2h} = \sum_{t=1}^{m_1} \frac{1}{2} \ln(s_t^2 + s_{ht}^2) + \sum_{t=1}^{m_2} DK_{ht} \,. \qquad (8)$$

Pokud by objekty byly charakterizovány pouze kvalitativními proměnnými, ze všech výše uvedených vzorců by byla vypuštěna ta část, která měří variabilitu kvantitativních proměnných, tj. výběrový rozptyl.

2 Měření podobnosti shluků

Podobnost shluků se zjišťuje například u aglomerativního hierarchického shlukování při postupném spojování nejpodobnějších shluků pro vytváření menšího počtu shluků.

Mezi koeficienty, které vyjadřují vzájemný vztah mezi objekty a shluky v případě, že jsou objekty charakterizované proměnnými různých typů, je možné zařadit **věrohodnostní míru**. Tato míra je využívána ve spojení s dvoukrokovou shlukovou analýzou v systému SPSS.

Při měření vzdálenosti D dvou shluků C_h a $C_{h'}$, které jsou charakterizovány současně pomocí kvantitativních a nominálních proměnných, se v tomto případě využívá entropie v kombinaci s výběrovým rozptylem a postupuje se tak, že se od hodnoty variability shluku vzniklého

spojením dvou shluků $H_{hh'}$ odečte součet hodnot variabilit těchto dvou samostatných shluků, tj.

$$D_{\mathrm{VMH}}(C_h, C_{h'}) = H_{hh'} - (H_h + H_{h'}) . \qquad (9)$$

V případě, že jsou objekty charakterizovány současně pomocí kvantitativních a nominálních proměnných je možné navrhnout alternativu k postupu (9), tj. měřit variabilitu pomocí kombinace Giniho koeficientu a výběrového rozptylu, a tedy postupovat podle vztahu

$$D_{\mathrm{VMG}}(C_h, C_{h'}) = G_{hh'} - (G_h + G_{h'}) . \qquad (10)$$

V případě, že jsou objekty charakterizovány současně pomocí kvantitativních a ordinálních proměnných je možné analogicky navrhnout měření variability pomocí kombinace koeficientu dorvar a výběrového rozptylu, tj. postupovat podle vztahu

$$D_{\mathrm{VDK}}(C_h, C_{h'}) = DK_{hh'} - (DK_h + DK_{h'}) . \qquad (11)$$

Závěr

Při vyjadřování podobnosti objektů pro případ, že jsou objekty charakterizovány pouze kvantitativními proměnnými existuje v současné literatuře řada měr. Pro případ, že jsou objekty charakterizovány vícehodnotovými kvalitativními proměnnými, existují k měření podobnosti pouze omezené možnosti. Vychází se z myšlenky, že dva objekty jsou si nejpodobnější, pokud shluk z nich vytvořený má nejmenší variabilitu. K jejímu měření se v praxi využívá entropie. Nově navrženým způsobem je měřit variabilitu pomocí Giniho koeficientu (v případě nominálních proměnných) či pomocí koeficientu dorvar, založeného na kumulativních relativních četnostech (v případě ordinálních proměnných). I v případě, že jsou objekty charakterizovány proměnnými různých typů, se vychází z myšlenky, že dva objekty jsou si nejpodobnější, pokud shluk z nich vytvořený má nejmenší variabilitu. K měření variability se v praxi používá výběrový rozptyl v kombinaci s entropií. Novým návrhem je použít pro měření variability rozptyl v kombinaci s hodnotou Giniho koeficientu či hodnotou koeficientu dorvar.

Při vyjadřování podobnosti shluků obsahující objekty, které jsou charakterizované kvalitativními proměnnými, se postupuje tak, že se od hodnoty variability shluku vzniklého spojením dvou shluků odečte součet hodnot variabilit těchto dvou samostatných shluků. V praxi se k tomu využívá entropie. Novým návrhem měření variability shluků je použití také hodnot Giniho koeficientu. Při vyjadřování podobnosti shluků obsahující objekty, které jsou

charakterizované proměnnými různých typů, se také postupuje tak, že se od hodnoty variability shluku vzniklého spojením dvou shluků odečte součet hodnot variabilit těchto dvou samostatných shluků. V praxi je variabilita shluků hodnocena pomocí měr s využitím výběrového rozptylu a entropie. Novým návrhem je použít pro měření variability shluků také kombinaci výběrového rozptylu a Giniho koeficientu pro nominální porměnné. Při praktických úlohách se na vybraných souborech ukázalo, že použití Ginniho koeficientu při shlukování je vhodnější, než v praxi používaná entropie, viz [4].

Literatura

[1] GAN, G., MA CH., WU J.: *Data Clustering Theory, Algorithms, and Applications*, ASA, Philadelphia, 2007.

[2] HALKIDI, M., BATISTAKIS, Y., VAZIRGIANNIS, M.: *Clustering algorithms and validity measures.* SSDBM, Athens, 2001.

[3] ŘEHÁK, J., ŘEHÁKOVÁ, B.: *Analýza kategorizovaných dat v sociologii*, Academia, Praha, 1986.

[4] ŘEZANKOVÁ, H., HÚSEK, D., LÖSTER, R.: *Clustering with Mixed Type Variables and Determination of Cluster Numbers*, CNAM and INRIA, Paříž, 2010, s. 1525-1532.

[5] ŘEZANKOVÁ, H., HÚSEK, D., SNÁŠEL, V.: *Shluková analýza dat,* 2. vydání, Professional Publishing, Praha, 2009.

[6] ŘEZANKOVÁ, H., HÚSEK, D.: Methods for the determination of the number of clusters in statistical software packages, VŠE KSTP; VŠE KMIE, Praha, 2008, s. 1-6.

[7] ŘEZANKOVÁ, H., LÖSTER, T., HÚSEK, D.: Evaluation of Categorical Data Clustering. Fribourg 26.01.2011 – 28.01.2011. In: Advances in Intelligent Web Mastering – 3. Berlin : Springer Verlag, 2011, s. 173–182.

Kontakt

Tomáš Löster, Ing., Ph. D.

Katedra statistiky a pravděpodobnosti

Fakulta informatiky a statistiky

Vysoká škola ekonomická v Praze

Nám. W. Churchilla 4, 130 67 Praha 3

Česká republika

Tel.: +420 2 24095 484

E-mail: tomas.loster@vse.cz