Podobnost objektů charakterizovaných nominálními proměnnými
Similarity of objects characterized by nominal variables

Řezanková Hana, Mohammad Adam

Základem některých metod vícerozměrné analýzy dat, jako jsou metody shlukové analýzy či vícerozměrného škálování, je stanovení míry podobnosti (příp. odlišnosti) objektů (příp. proměnných či kategorií nominální proměnné). V případě, kdy jsou objekty charakterizovány kvantitativními proměnnými, jsou velmi dobře známy a používány míry (jak vzdálenosti, tak podobnosti), které jsou běžně implementovány ve statistických programových systémech (euklidovská vzdálenost a další). Obsahuje-li datový soubor proměnné jiného typu, případně různých typů, má analytik podstatně omezenější možnosti, i když v literatuře již byla navržena celá řada měr, případně jiné metody shlukování využitelné pro takové případy.

To je důvodem, proč jsou dříve navržené míry podobnosti objektů charakterizovaných nominálními proměnnými stále hodnoceny, porovnávány a navrhovány nové přístupy ke stanovování míry podobnosti. Příznivá situace je v případě dichotomických proměnných, pro který jednak existuje značné množství speciálních měr podobnosti a odlišnosti (byť jejich zařazení do programových systémů není samozřejmostí), jednak v této situaci mohou být využity (a jsou využívány) míry pro kvantitativní proměnné. Úkolem analytika je „pouze“ převést každou nominální proměnné s více než dvěma kategoriemi na skupinu pomocných (indikátorových) binárních proměnných, které indikují výskyt jednotlivých kategorií. Je tak vytvořen soubor binárních dat, který lze snadno analyzovat standardními prostředky.

Existují ovšem i jiné přístupy, kterým je v literatuře stále věnována určitá pozornost. V článku [6] se sice vychází z pomocného souboru binárních dat, jsou však zohledněny skupiny indikátorových proměnných vyjadřující přítomnost kategorií původních proměnných. Jsou navrženy váhy jednak pro jednotlivé indikátorové proměnné (základem je relativní četnost jedniček), jednak pro jejich skupiny odpovídající původním proměnným. Odvozen je i postup stanovení míry podobnosti v případě, kdy jsou proměnné závislé.

Četnosti výskytu jedniček jsou zohledněny u několik měr podobnosti, navržených již v 60. a 70. letech minulého století, ale i později. V literatuře jsou buď hledány způsoby, jak tyto míry hodnotit a porovnávat, viz [1], případně jsou navrhovány nové míry či postupy a porovnávány s těmito staršími, viz [3] až [5]. Jeden z navržených postupů nejprve stanovuje shluky proměnných na základě jejich zjištěných vzdáleností a poté vyjadřuje charakteristiky skupin objektů, viz [2].

Cílem tohoto příspěvku je zhodnotit vybrané míry podobnosti, které byly navrženy pro objekty charakterizované nominálními proměnnými, a to na ilustračním příkladu při rozdělení původní množiny objektů do různých počtů shluků pomocí hierarchické shlukové analýzy při využití spojování pomocí metody nejvzdálenějšího souseda. Hodnocení shlukování je provedeno na základě postupů navržených v článku [7]. Slouží k němu vnitroshluková variabilita vyjádřená buď normovanou Giniho mírou, nebo normovanou entropií, analogie Goodmanova-Kruskalova tau a koeficientu nejistoty pro více vysvětlovaných proměnných a analogie pseudo F indexu opět s použitím buď Giniho míry, nebo entropie.

Literatura

[1] BORIAH, S. – CHANDOLA, V. – KUMAR, V. 2008. Similarity measures for categorical data: A comparative evaluation. In: Proceedings of the 8th SIAM International Conference on Data Mining. 2008, s. 243–254.

[2] DAS, G. – MANNILA, H. 2000. Context-based similarity measures for categorical databases. In: Principles of Data Mining and Knowledge Discovery. Lecture Notes in Computer Science, Vol. 1910, 2000, s. 201–210.

[3] DESAI, A. – SINGH, H. – PUDI, V. 2011. DISC: Data-intensive similarity measure for categorical data. In: Advances in Knowledge Discovery and Data Mining. Lecture Notes in Computer Science, Vol. 6635, 2011, s. 469–481.

[4] CHANDOLA, V. – BORIAH, S. – KUMAR, V. 2009. A framework for exploring categorical data. In: Proceedings of the 9th SIAM International Conference on Data Mining, 2009, s. 187–198.

[5] LE, S.Q. – HO, T.B. 2005. An association-based dissimilarity measure for categorical data. In: Pattern Recognition Letters, roč. 26, 2005, s. 2549–2557.

[6] MORLINI, I. – SERGIO, Z. 2012. A new class of weighted similarity indices using polytomous variables. In: Journal of Classification, roč. 29, 2012, s. 199–226.

[7] ŘEZANKOVÁ, H., LÖSTER, T., HÚSEK, D. 2011. Evaluation of categorical data clustering. In: Advances in Intelligent Web Mastering – 3. Berlin : Springer Verlag, 2011, s. 173–182.

Adresa autorů:

Řezanková Hana, prof. Ing., CSc.
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
hana.rezankova@vse.cz

Mohammad Adam
Vysoká škola ekonomická v Praze
nám. W. Churchilla 4, 130 67 Praha 3
xmoha00@vse.cz