Modification of CHF and BIC coefficients for Evaluation of Clustering with Mixed Type Variables

Tomas Löster *
University of Economics, Prague, Czech Republic, tomas.loster@vse.cz

Tomas Pavelka
University of Economics, Prague, Czech Republic, pavelkat@vse.cz

Current literature draws attention particularly to the evaluation of clustering in a situation when individual objects are characterized only by quantitative variables. The problems associated with the analysis of data characterized by qualitative or mixed type variables have only been dealt with to a limited extent. This is based on an analogy of the techniques applied when evaluating log-linear models for example. In this paper we suggest new coefficients for the evaluation of resulting clusters based on the principle of the variability analysis. Furthermore, only coefficients for mixed type variables based on a combination of sample variance and one of the variability measures for nominal variables will be presented. Similar approaches can be applied in the case of qualitative variables while omitting the part characterizing the variability of quantitative variables. In this paper we evaluated selected indices for determining the number of clusters when objects are characterized by mixed type variables too. On the basis of real data files analyses (Database The UCI Machine Learning Repository website: http://archive.ics.uci.edu/ml/datasets.html) we compared three newly proposed indices with the known BIC criterion, which is implemented in two-step cluster analysis in the IBM SPSS Statistics system. We knew the number of object groups and We were interested in agreement of the found optimal number of clusters with the real number of groups. We had analyzed 33 data files and it was found that new indices determined the correct number of clusters more successful than BIC criterion which is implemented in two-step cluster analysis in the IBM SPSS Statistics system. Criterions based on Gini coefficient were more successful than criterion based on Entropy. The CHFG index determined the correct number of clusters in most cases (93,33 %). The second successful criterion was the CHFH index (73,33 %). The BIC criterion determines the correct number of clusters in 40,0 % of cases and our modification of BIC criterion (using Gini coefficient insted of entropy, which is used in known BIC criterion) was successful in 46,67 % of cases.

Key Words: Cluster analysis, evaluation of Clustering, BIC criterion, CHF criterion