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

Tomas Löster *
University of Economics, Prague, Czech Republic,
Faculty of Informatics and Statistics,
Department of Statistics and Probability,
email: tomas.loster@vse.cz

Tomas Pavelka
University of Economics, Prague, Czech Republic,
Faculty of Business Administration,
Department of Microeconomics,
email: pavelkat@vse.cz

Abstract

Cluster analysis is a multivariate statistical method, which is used to classify objects. It is used in many areas, such as the classification of customers or respondents in various marketing surveys. Individual objects are characterized by different variables. Variables can be quantitative and qualitative. Depending on the type of variables it is necessary to select the appropriate method of measuring distances of objects and clusters. There are many ways how to measure these distances and it is not clearly defined how to choose specific measure in different conditions. Depending on the extent of distances and the method chosen may arise different clusters, and thus different results. For this reason, it is necessary to evaluate the clustering result. The evaluation should analyze the numbers of clusters and different clustering methods. There are many coefficients for evaluate results of clustering. In the current literature are defined in particular coefficients, which are used for the quantitative variables. For variables of mixed types (a combination of qualitative and quantitative) are coefficients described only in a very limited extent. The aim of this paper is to analyze the modified coefficients CHF and BIC on real data sets in case of mixed types variables.

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

1. Introduction

Cluster analysis is a multivariate statistical method, which is used to classify objects. It is used in many areas, such as the classification of customers or respondents in various marketing surveys. Cluster analysis involves a broad scale of techniques. Hence an important factor when examining data structure is therefore the comparison of resulting clusters obtained by various algorithms and selection of the best assignment of objects to clusters. Determining the optimal number of clusters is also important.

Current literature draws attention particularly to the evaluation of clustering in a situation when individual objects are characterized only by quantitative variables,see Gan (2007), Halkidi (2001).

The aim of this paper is to analyze the modified coefficients CHF and BIC on real data sets in case of mixed types variables.

For determining the number of clusters we suggest to modify *the CHF index*, which is defined in Gan (2007). We modify *the CHFH index* in the form

$$I_{\mathrm{CHFH}}(k) = \frac{(n-k)\cdot\left[H(1)-H(k)\right]}{(k-1)\cdot H(k)}, \qquad (1)$$

or *the CHFG index* in the form

$$I_{\mathrm{CHFG}}(k) = \frac{(n-k)\cdot\left[G(1)-G(k)\right]}{(k-1)\cdot G(k)}, \qquad (2)$$

where n is the number of objects, k is number of clusters, and

$$H(k) = \sum_{h=1}^{k}\frac{n_h}{n}\left(\sum_{t=1}^{m_1}\frac{1}{2}\ln(s_t^2+s_{ht}^2)+\sum_{t=1}^{m_2}\left(-\sum_{u=1}^{K_t}\left(\frac{n_{htu}}{n_h}\ln\frac{n_{htu}}{n_h}\right)\right)\right), \qquad (3)$$

where m_1 is the number of quantitative variables, m_2 is the number of nominal variables, s_t^2 is the sample variance of the tth variable, s_{ht}^2 is the sample variance of the tth variable in the hth cluster, K_t is the number of categories of the tth variable, n_{htu} is the frequency of the uth category of the tth variable in the hth cluster, and n_h is the number of objects in the hth cluster, and where

$$G(k) = \sum_{h=1}^{k}\frac{n_h}{n}\left(\sum_{t=1}^{m_1}\frac{1}{2}\ln(s_t^2+s_{ht}^2)+\sum_{t=1}^{m_2}\left(1-\sum_{u=1}^{K_t}\left(\frac{n_{htu}}{n_h}\right)^2\right)\right). \qquad (4)$$

The high values of I_{CHFH} or I_{CHFG} indicate well separated clusters, i.e. the maximum value within a certain interval is searched.

The *Schwarz Bayesian information criterion* (BIC) can also be applied to determine the optimal number of clusters, see Řezanková (2010). It can be calculated according to the formula

$$I_{\mathrm{BIC}}(k) = 2H(k) + k(2m_1 + \sum_{t=1}^{m_2}(K_t-1)\ln(n)\,. \qquad (5)$$

We newly suggest also used $G(k)$ instead of $H(k)$. This criterion will be denoted as I_{BICG} in the following text and it can be calculated according to the formula

$$I_{\mathrm{BICG}}(k) = 2G(k) + k(2m_1 + \sum_{t=1}^{m_2}(K_t-1)\ln(n)\,. \qquad (6)$$

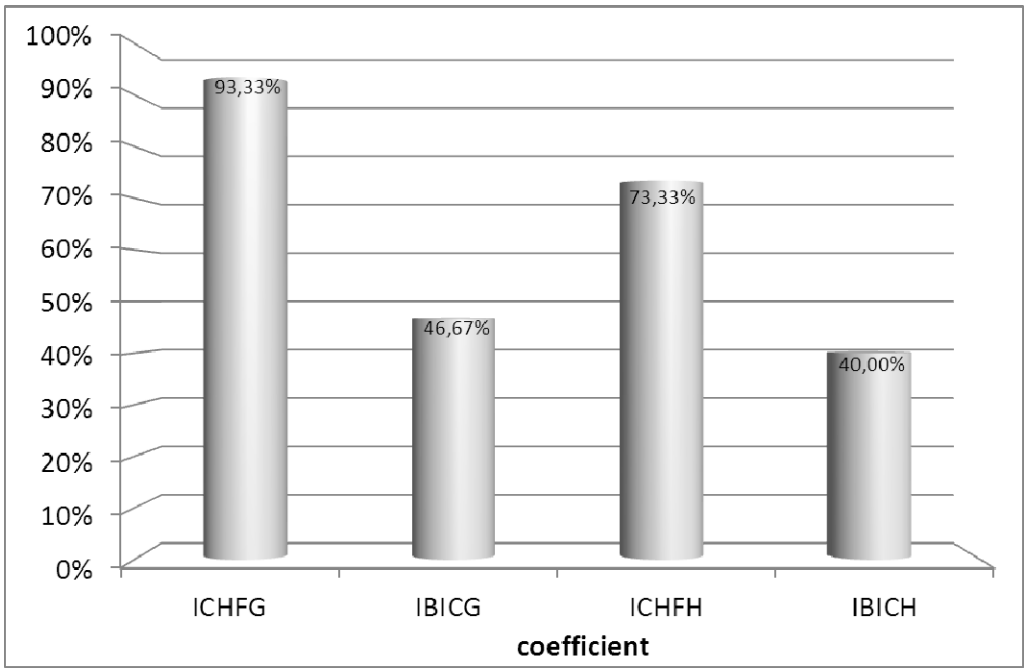
The estimate of the number of clusters is determined on the basis of the minimum value of this coefficient.

2. Results

In this part of our paper we discuss the results and conclusions of the practical application of suggested coefficients applicable to mixed type variables. We used data files from the UCI Machine Learning Repository, see http://archive.ics.uci.edu/ml/datasets.html) are analyzed. In all cases we used two cluster analysis, which is implemented to the IBM SPSS system. The BIC index is stated as a representant of the existing coefficients for a comparison with newly proposed indices.

We used totally 33 data files, for example Wine File, the German credit data file, IRIS from the UCI Machine Learning Repository etc. We knew correct number of clusters. We used results from SPSS system – we received membership of objects to clusters and I calculated all modified criterions.

Figure 1: Success rate of individual criterions in given of number of clusters



Source: Own calculation

As we can see from figure 1, 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 % cases and my modification of BIC criterion (using Gini coefficient instead of Entropy, which is used in known BIC criterion) was successful in 46.67 % of cases.

3. Conclusion

In this paper we evaluated selected indices for determining the number of clusters when objects are characterized by mixed type variables. 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. Criterion based on Gini coefficient (I_{CHFG} and I_{BICG}) were more successful than criterion based on Entropy (I_{CHFH} and I_{BIC}).
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 % cases and my modification of BIC criterion (using Gini coefficient instead of Entropy, which is used in known BIC criterion) was successful in 46.67 % of cases.
In conclusion, we can say that our modifications are more successful in evaluating the results of clustering than the commonly used coefficient BIC, which is implemented in the IBM SPSS. We can say, that modifications using the Gini coefficient are also more successful than modifications that uses entropy too.

Acknowledgements: This article was created with the help of the Internal Grant Agency of University of Economics in Prague MF/F4/6/2013.

References

Calinski, T., Harabasz, J.: A dendrite method for cluster analysis, *Comunications in Statistics*, Vol. 3, 1974, 1–27.

Gan, G., Ma, C., Wu, J.: *Data Clustering Theory, Algorithms, and Applications*. ASA, Philadelphia, 2007.

Halkidi, M., Batistakis, Y., Vazirgiannis, M.: *Clustering Algorithms and Validity Measures.* SSDBM, Athens, 2001.

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

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

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

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.

UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/datasets.html