VOLUME 8, ISSUE 2, 2013

ISSN 2045-3345

ISSN 2045-3345

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

RESEARCH JOURNAL OF ECONOMICS, BUSINESS AND ICT

MODIFICATION OF CHF AND BIC COEFFICIENTS FOR EVALUATION OF CLUSTERING WITH MIXED TYPE VARIABLES

Tomáš Löster[I]
University of Economics, Prague, Czech Republic

ABSTRACT

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 I 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 I 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) I compared three newly proposed indices with the known BIC criterion, which is is implemented in two-step cluster analysis in the IBM SPSS Statistics system. I knew the number of object groups and I was interested in agreement of the found optimal number of clusters with the real number of groups. I had analyzed 15 data files and it was found that new indices determined the correct number of clusters more successful than BIC criterion which is 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 my modification of BIC criterion (using Gini coefficient insted of entropy, which is used in known BIC criterion) was successful in 46.67 % of cases.

JEL CLASSIFICATION & KEYWORDS

■ C38 ■ C40 ■ CLUSTER ANALYSIS ■ EVALUATION OF CLUSTERING ■ BIC CRITERION ■ CHF CRITERION

INTRODUCTION

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.

[I] tomas.loster@vse.cz

Current literature draws attention particularly to the evaluation of clustering in a situation when individual objects are characterized only by quantitative variables, see [2], [3]. 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 I 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.

The following text is organized in such a way that in Section 2 there is a description of the newly proposed coefficients and in Section 3 these coefficients are applied for determining the optimal number of clusters in real data files. Conclusion presents an evaluation of the obtained findings.

Evaluation of clustering results in case of mixed type variables

In this paper disjunctive clustering resulting in the unique assignment of objects to clusters is only considered. If objects are characterized only by qualitative variables it can be accomplished, for example, using hierarchical cluster analysis with the application of the coefficient of disagreement as a dissimilarity measure, see [5]. In case of mixed type variables a log-likelihood distance measure can be applied (it is implemented in two-step cluster analysis in the IBM SPSS Statistics system, see [7]).

The evaluation of the results of clustering can be based on within-cluster variability. The method is better which results in clusters with less variability. To determine variability in case that objects are characterized by mixed type variables, a combination of sample variance and entropy, which is defined in [4], is applied in practice (in the SPSS system). Within-cluster variability for k clusters is determined by the formula

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

where n is the number of objects, m_1 is the number of quantitative variables, m_2 is the number of nominal variables, s^2_t is the sample variance of the tth variable, s^2_{ht} 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. I have proposed several coefficients for clustering evaluation both for the analysis with categorical variables and for mixed type variables.

As an alternative to Formula (1) I suggest a measure which applies a combination of the sample variance and the Gini coefficient. For *k* clusters it can be determined according to the formula

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

For determining the number of clusters I suggest to modify the CHF index, which is defined in [1]. I can use either Formula (1) or Formula (2) as a variability measure, i.e. I obtain either the CHFH index in the form

$$I_{\mathrm{CHFH}}(k)=\frac{(n-k)\big(H(1)-H(k)\big)}{(k-1)H(k)} \quad (3)$$

or the CHFG index in the form

$$I_{\mathrm{CHFG}}(k)=\frac{(n-k)\big(G(1)-G(k)\big)}{(k-1)G(k)} \quad (4)$$

The high values of these indices 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 [6]. It can be calculated according to the formula

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

I 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

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

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

Application of new indices of evaluation to real data files

This part describes the results and conclusions of the practical application of the newly suggested coefficients applicable to mixed type variables. Data files from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets.html) are analyzed. The BIC index is stated as a representant of the existing coefficients for a comparison with newly proposed indices.

Some data files were adapted for mixed typed variables, because there is not large number of real data files (in the UCI Machine Learning Repository, etc.) for clustering with mixed type of variables in case that real (optimal) number of clusters is known. For example I created categorical variable from quantitative variable. These data files are called with “number” of variant.

There are examples of results for evaluating in the Wine File and the German credit data file in next paragraphs. In all cases I used number of clusters from interval 1 – 15, because I knew correct (real) number of clusters, which was no higher than 15. In all cases I used two-set cluster analysis for clustering of objects (in the IBM SPSS system), because it is method for clustering with mixed type variables, which is implemented to commonly using statistical software. Methods for clustering with mixed type of variables are very limited in common using software. From SPSS system I received membership of objects to clusters (for all cases from interval 1 – 15) and I calculated all criterions.

The wine file

The Wine file includes 178 wine samples. The original data file contains thirteen quantitative variables which express the the quantities of constituents. I created categories for two variables (Flavanoids and Prolines) in order to analyze the file with mixed type variables. For each of the wines the classification into some of three groups representing different cultivars is known.

Analysis of clustering results – Variant 1

In this part I present results of the analysis of the file with eleven quantitative variables and two recoded three-category variables. Table 1 shows the values of indices described in this paper. Three clusters which correspond to the correct number of groups were found as optimal on the basis of all indices.

Table 1: Evaluation of Variant 1

k	$I_{\mathrm{BIC}}(k)$	$I_{\mathrm{CHFH}}(k)$	$I_{\mathrm{BICG}}(k)$	$I_{\mathrm{CHFG}}(k)$
2	1896.95	55.88	1698.18	50.41
3	**1697.34**	**57.59**	**1555.43**	**52.19**
4	1730.45	46.37	1618.77	40.71
5	1788.90	39.90	1707.69	33.62
6	1867.52	35.24	1805.60	29.00
7	1945.47	32.47	1899.36	26.28
8	2035.82	30.7	1999.44	24.14
9	2134.30	28.2	2099.98	22.63
10	2239.16	26.21	2215.27	20.86
11	2350.72	24.52	2331.25	19.44
12	2458.29	23.36	2439.57	18.62
13	2567.92	22.36	2549.2	17.93
14	2678.14	21.54	2664.78	17.16
15	2798.98	20.44	2787.96	16.25

Source: Own calculation

Analysis of clustering results – Variant 2

In the second variant I analyzed the file with eleven quantitative variables and two recoded four-category variables. Table 2 shows the values of indices described in this paper. Four clusters were selected as optimal according to the known BIC criterion. On the basis of CHFH index two clusters were selected as optimal. It is therefore obvious that in these cases the correct number of clusters has not been determined. According to the BICG and CHFG indices (using a combination of sample variance and the Gini coefficient and) three clusters were found as optimal, and thus the correct number was found.

When analysing this file, it was therefore found that the newly suggested indices based on a combination with

Table 2: Evaluation of Variant 2

k	IBIC(k)	ICHFH(k)	IBICG(k)	ICHFG(k)
2	2080.05	**52.90**	1834.91	39.01
3	1916.43	50.02	**1717.62**	**41.26**
4	**1912.55**	43.35	1743.15	36.13
5	1974.09	37.38	1842.76	29.80
6	2046.29	33.71	1929.39	26.91
7	2127.80	31.15	2033.13	24.36
8	2224.48	28.86	2142.56	22.39
9	2319.05	27.41	2259.34	20.68
10	2421.55	26.10	2373.33	19.52
11	2529.41	24.95	2490.09	18.55
12	2642.27	23.89	2615.63	17.47
13	2768.76	22.52	2740.57	16.62
14	2888.93	21.63	2864.52	15.96
15	3016.01	20.64	2995.29	15.19

Source: Own calculation

the Gini coefficient can better determine the number of clusters than indices based on entropy.

The German credit data file

The German Credit Data file (the Statlog name is also cited) includes 1.000 objects (customers). The file contains seven quantitative variables (e.g. age in years, credit amount) and thirteen qualitative variables (e.g. personal status and sex, type of housing). For each of the customers the classification into some of two groups representing different level of risk is known.

I analyzed the file with all variables. In Tables 3 there are values of all investigated indices. According to all indices two clusters were determined as optimal, which is the correct number.

Table 3: Evaluation of Credit Data Clustering

k	$I_{BIC}(k)$	$I_{CHFH}(k)$	$I_{BICG}(k)$	$I_{CHFG}(k)$
2	**22980.8**	**90.26**	**15418.4**	**75.85**
3	23085.3	69.37	15811.7	63.37
4	23357.7	60.45	16376.7	55.41
5	23669.3	56.17	17001.7	50.63
6	24137.9	52.05	17643.2	47.87
7	24739.2	48.05	18427.0	43.77
8	25371.6	45.03	19188.6	41.33
9	26059.6	42.37	20031.4	38.43
10	26800.7	39.91	20896.2	35.94
11	27561.6	37.85	21733.6	34.32
12	28372.8	35.82	22641.9	32.24
13	29160.1	34.33	23517.7	30.86
14	29930.5	33.22	24381.8	29.84
15	30769.2	31.86	25308.3	28.40

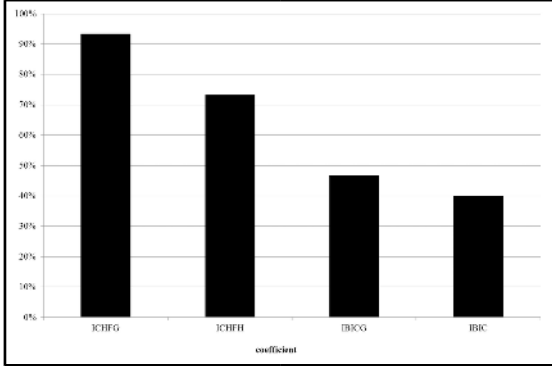
Source: Own calculation

Conclusion of results of all analyzed data files are placed in table 4. Correct numbers of clusters in individual data files are placed in second column. Other columns contain evaluation of given results of individual criterion.

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 insted of Entropy, which is used in known BIC criterion) was successful in 46.67 % of cases.

Success rate of individual criterions in all analyzed data files is obvious from figure 1.

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



Source: Own calculation

Conclusion

In this paper I suggested and 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) I compared three newly proposed indices with the known BIC criterion, which is implemented in two-step cluster analysis in the IBM SPSS Statistics system. I knew the number of object groups and I was interested in agreement of the found optimal number of clusters with the real number of groups. I had analyzed 15 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 systém. Criterion 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

Table 4: Summarization of results of individual criterions

File/criterion	***correct number of clusters***	I_{BIC}	I_{BICG}	I_{CHFH}	I_{CHFG}
Car Evaluation	4	incorrect	incorrect	**correct**	**correct**
Adult	2	incorrect	incorrect	**correct**	**correct**
Wine 1	3	**correct**	**correct**	**correct**	**correct**
Wine 2	3	incorrect	**correct**	incorrect	**correct**
Wine 3	3	**correct**	**correct**	incorrect	**correct**
Wine 4	3	**correct**	**correct**	**correct**	**correct**
Wine 5	3	**correct**	**correct**	**correct**	**correct**
Iris 1	3	incorrect	**correct**	**correct**	**correct**
Iris 2	3	incorrect	incorrect	**correct**	**correct**
Contraceptive 1	3	incorrect	incorrect	**correct**	**correct**
Contraceptive 2	3	incorrect	incorrect	incorrect	incorrect
Cardiotocography	10	incorrect	incorrect	correct	**correct**
Thyroid Disease	6	incorrect	incorrect	incorrect	**correct**
German Credit 1	2	**correct**	**correct**	**correct**	**correct**
German Credit 2	2	**correct**	**correct**	**correct**	**correct**
Ratio of successfully given number of clusters	–	**40.00%**	**46.67%**	**73.33%**	**93.33%**

Source: Own calculation

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.

Acknowledgements

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

References

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

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

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

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

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

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

[7] 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.

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