Metodološki zvezki, Vol. 15, No. 2, 2018, 1–20

Internal Evaluation Criteria for Categorical Data in Hierarchical Clustering: Optimal Number of Clusters Determination

Zdeněk Šulc[1] Jana Cibulková[2] Jiří Procházka[3]
Hana Řezanková[4]

Abstract

The paper compares 11 internal evaluation criteria for hierarchical clustering of categorical data regarding a correct number of clusters determination. The criteria are divided into three groups based on a way of treating the cluster quality. The variability-based criteria use the within-cluster variability, the likelihood-based criteria maximize the likelihood function, and the distance-based criteria use distances within and between clusters. The aim is to determine which evaluation criteria perform well and under what conditions. Different analysis settings, such as the used method of hierarchical clustering, and various dataset properties, such as the number of variables or the minimal between-cluster distances, are examined. The experiment is conducted on 810 generated datasets, where the evaluation criteria are assessed regarding the optimal number of clusters determination and mean absolute errors. The results indicate that the likelihood-based BIC1 and variability-based BK criteria perform relatively well in determining the optimal number of clusters and that some criteria, usually the distance-based ones, should be avoided.

1 Introduction

Cluster analysis is a multivariate statistical method that reveals an underlying structure of data by identifying homogeneous groups (clusters) of objects. The homogeneity is defined as possessing a certain relevant property by the majority of objects in the group (de Souto et al., 2012). Under the term cluster analysis, many clustering algorithms are included, each of them with several possible similarity measures. Different algorithms can lead to different object assignments, and thus, a comparison of several object assignments

[1]Department of Statistics and Probability, University of Economics, Prague, Czech Republic; zdenek.sulc@vse.cz
[2]Department of Statistics and Probability, University of Economics, Prague, Czech Republic; jana.cibulkova@vse.cz
[3]Department of Statistics and Probability, University of Economics, Prague, Czech Republic; jiri.prochazka@vse.cz
[4]Department of Statistics and Probability, University of Economics, Prague, Czech Republic; hana.rezankova@vse.cz

using one or more evaluation criteria is often welcomed. For cluster partition evaluation, external or internal evaluation criteria are commonly used. The external criteria, see, e.g., de Souto et al. (2012), are based on comparing a cluster assignment to an a priori-known class variable. Apart from the simulation studies, where the values of a class variable are known, they are not suitable for clustering evaluation. The internal criteria, see, e.g., Liu et al. (2010), Milligan and Cooper (1985), Vendramin et al. (2010), use intrinsic properties of a dataset. Hence, they are more suitable for the unsupervised methods. Moreover, the internal criteria can be further divided either to the criteria trying to determine the optimal number of clusters or to judge the quality of a particular cluster solution, see Arbelaitz et al. (2013). Some of the criteria were developed for both these tasks.

Studies, some of which have been mentioned in the previous paragraph, deal with evaluation criteria for quantitative data, and the majority of them cannot be used for categorical data. There are two main reasons for that. First, many evaluation criteria for quantitative data use mathematical operations between the values of the raw data matrix. That is not possible if the categorical values are used. Thus, the evaluation criteria for categorical data can be only based on calculations within a dissimilarity matrix, which is numeric also for categorical data. Second, the commonly used concepts for cluster evaluation used in quantitative data, such as the variance to express variability, cannot be used directly, but they have to be adjusted for their use in categorical data or appropriately substituted by a categorical data alternative.

The clustering of categorical data is not as trustworthy as the quantitative one. The reason is the lower variability of categories by categorical variables compared to that by quantitative variables, which does not enable distinguishing groups in the data so precisely. However, there are many situations when clustering of purely categorical data is necessary (medicine, psychology, marketing), and for such cases, one should have a few reliable criteria to evaluate the obtained clusters. There is a lack of papers comparing and assessing the evaluation criteria determined for categorical data. Although there are many papers that use the internal evaluation criteria for categorical data, see Bontemps and Toussile (2013), where the BIC and AIC criteria for categorical data are used, or Řezanková et al. (2011), where the variability-based criteria are applied, we found none which compares internal evaluation criteria for categorical data. Thus, this paper tries to fill this gap by presenting and comparing different approaches which researchers can use to evaluate their categorical clustering outputs.

This paper compares selected internal evaluation criteria for categorical data, which are determined for the optimal number of clusters determination in hierarchical cluster analysis (HCA). The criteria are evaluated on different cluster analysis settings (similarity measures, linkage methods) and also on the generated dataset properties (number of variables, categories, and clusters) regarding their ability to identify the optimal number of clusters. This can be formulated as two main aims. The first one is to evaluate the performance of the examined evaluation criteria regarding the correct number of clusters determination. The second one is to determine which properties of the clustered datasets associate with the outcomes of the examined evaluation criteria. To achieve this aim, logistic regression is used. In the experiment, 11 internal evaluation criteria are compared and assessed. The criteria are grouped according to the principles they use: variability, likelihood, and distance. The variability-based criteria express the cluster quality by their low within-cluster variability, the likelihood-based criteria assess it by the low values of

the likelihood function approximation for the categorical data, and the distance-based criteria express it by the low within-cluster distances. The experiment is performed on 810 generated datasets with known cluster assignments and certain properties under control, such as the numbers of clusters or variables.

Paper is organized as follows. Section 2 presents the examined internal evaluation criteria for categorical data. Section 3 focuses on the selected similarity measures and methods of HCA. Section 4 describes the dataset generation process and the experiment settings. The results are presented in Section 5, and the outcomes of the research are summarized in the Conclusion.

2 Internal Evaluation Criteria

Since the clusters in a dataset should be ideally distinct and their objects similar, internal evaluation criteria are usually constructed with the aim to satisfy the assumptions of *compactness* and *separation* of the created clusters (Liu et al., 2010; Zhao et al., 2002). Whereas the compactness measures the similarity of the objects in clusters, the separation measures distinctness between the clusters. The evaluation criteria presented in this section measure the compactness and separation based on the principles of either a variability, a likelihood or a distance.

2.1 Evaluation Criteria Based on the Variability

The variability-based evaluation criteria are usually based on the compactness principle, which is expressed by the low within-cluster variability of the created clusters. In this subsection, three internal evaluation criteria based on this principle are presented, because they performed well in (Šulc, 2016) and (Yang, 2012), namely the *pseudo F index based on the mutability* (PSFM), the *pseudo F index based on the entropy* (PSFE) and the BK index.

The PSFM index (Řezanková et al., 2011) is based on the within-cluster variability expressed by the mutability (the Gini coefficient), see Gini (1912) that appears in Light and Margolin (1971). It can be written as

$$PSFM(k) = \frac{(n-k)\,(WCM(1) - WCM(k))}{(k-1)\,WCM(k)}, \tag{2.1}$$

where k represents the total number of clusters and n is the number of objects in a dataset. $WCM(1)$ and $WCM(k)$ represent the within-cluster variability in the whole dataset and the k-cluster solution moving in a range from zero (no variability) to one (maximal variability). $WCM(k)$ is computed as

$$WCM(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} G_{gc},$$

where n_g is the number of objects in the g-th cluster ($g = 1, 2, \ldots, k$), m is the total number of variables and G_{gc} is the mutability by the c-th variable ($c = 1, 2, \ldots, m$) in the

g-th cluster expressed as

$$G_{gc} = 1 - \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \right)^2,$$

where n_{gcu} is the number of objects in the g-th cluster by the c-th variable with the u-th category ($u = 1, \ldots, K_c$) and K_c is the number of categories by the c-th variable.

The PSFE index (Řezanková et al., 2011) is constructed analogically to (2.1) with the difference that instead of $WCM(k)$, the variability $WCE(k)$ based on the entropy is used. $WCE(k)$ can be expressed as

$$WCE(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} H_{gc},$$

where H_{gc} is the entropy by the c-th variable in the g-th cluster according to the formula

$$H_{gc} = -\sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \ln \frac{n_{gcu}}{n_g} \right).$$

Both PSFM and PSFE indices indicate the optimal number of clusters by their maximal value over several examined cluster solutions. In such a cluster solution, the highest decrease in the within-cluster variability occurs.

The BK index (Chen and Liu, 2009) is defined as the second-order difference of the incremental entropy of the dataset with k clusters

$$BK(k) = \Delta^2 I(k) = (I(k-1) - I(k)) - (I(k) - I(k+1)),$$

where $I(k)$ is the incremental expected entropy in the k-cluster solution with the formula

$$I(k) = H_E(k) - H_E(k+1),$$

where H_E is the expected entropy in a dataset expressed as

$$H_E(k) = \sum_{g=1}^{k} \frac{n_g}{n} \sum_{c=1}^{m} \frac{H_{gc}}{\ln K_c}.$$

The highest value of the index indicates the optimal number of clusters.

2.2 Evaluation Criteria Based on the Likelihood

The likelihood-based evaluation criteria maximize the likelihood of the data while penalizing complex models (Biem, 2003). There are two commonly used evaluation criteria based on the likelihood for cluster quality assessment, the Bayesian information criterion (BIC) and the Akaike information criterion (AIC). In this paper, four of their modifications for categorical data based on the mutability and the entropy are presented. All of them indicate the optimal number of clusters by their minimal value.

The modification of the BIC index (Schwarz, 1978) for the categorical data using the entropy was presented in the SPSS manual (SPSS, Inc., 2001) and further described in

Bacher et al. (2004). Its calculation consists of two steps. In the first one, the modified index for categorical data is computed, and in the second one, the outputs provided in the first step are further refined. The first step can be written down as

$$BIC1(k) = -2\sum_{g=1}^{k} n_g \sum_{c=1}^{m} H_{gc} + k\sum_{c=1}^{m}(K_c - 1)\ln n.$$

A modification of this coefficient using the mutability with the formula

$$BIC2(k) = -2\sum_{g=1}^{k} n_g \sum_{c=1}^{m} G_{gc} + k\sum_{c=1}^{m}(K_c - 1)\ln n \tag{2.2}$$

was introduced in Löster (2013).

The modification of the AIC index (Akaike, 1973) for the categorical data using the entropy (SPSS, Inc., 2001; Bacher et al., 2004) can be expressed as

$$AIC1(k) = -2\sum_{g=1}^{k} n_g \sum_{c=1}^{m} H_{gc} + 2k\sum_{c=1}^{m}(K_c - 1). \tag{2.3}$$

Analogously to (2.2) and (2.3), a modification of the AIC index based on the mutability was derived in the same manner. It is expressed as

$$AIC2(k) = -2\sum_{g=1}^{k} n_g \sum_{c=1}^{m} G_{gc} + 2k\sum_{c=1}^{m}(K_c - 1).$$

For all four information criteria, the second step of the computation is the same. According to Bacher et al. (2004), it is defined as follows. First, a ratio $R(k)$ is defined as

$$R(k) = \frac{d_{k-1}^{(H)}}{d_k^{(H)}} \quad \text{or} \quad R(k) = \frac{d_{k-1}^{(G)}}{d_k^{(G)}},$$

where d_k represents a change of the entropy H or mutability G between a cluster solution containing $k-1$ clusters and k clusters. It is defined as a difference in entropies

$$d_k^{(H)} = H(k-1) - H(k)$$

or mutabilities

$$d_k^{(G)} = G(k-1) - G(k),$$

where $H(k)$ resp. $G(k)$ express the variability in the k-cluster solution; for instance, $H(k)$ is defined as:

$$H(k) = \sum_{g=1}^{k} n_g \sum_{c=1}^{m} H_{gc}.$$

Thus, the d_k statistic for the entropy is defined as

$$d_k^{(H)} = \sum_{h=1}^{k-1} n_h \sum_{c=1}^{m} H_{hc} - \sum_{g=1}^{k} n_g \sum_{c=1}^{m} H_{gc},$$

where h is the number of clusters ($h = 1, 2, \ldots, k-1$), and as

$$d_k^{(G)} = \sum_{h=1}^{k-1} n_h \sum_{c=1}^{m} G_{hc} - \sum_{g=1}^{k} n_g \sum_{c=1}^{m} G_{gc}$$

for criteria based on the mutability.

Next, for the two largest values of $R(k)$, $\max_1$ and $\max_2$, a ratio r is computed:

$$r = \frac{\max\limits_{k \geq 2}{}_1 R(k)}{\max\limits_{k \geq 2}{}_2 R(k)}.$$

If the ratio is higher than the threshold value $t = 1.15$, see Bacher et al. (2004), the number of clusters belonging to $\max\limits_{k \geq 2}{}_1 R(k)$ is chosen. Otherwise, the higher number of clusters from $\max\limits_{k \geq 2}{}_1 R(k)$ and $\max\limits_{k \geq 2}{}_2 R(k)$ is chosen.

2.3 Evaluation Criteria Based on the Distances

The distance-based criteria usually utilize both principles of compactness and separation. Satisfying the compactness principle, clusters should be of a small size, and satisfying the separation principle, their distance to the other clusters should be sufficiently high. The internal distance-based criteria from the NbClust R package (Charrad et al., 2014) were selected for comparison. In this package, there are 30 criteria of this type, but 25 of them require the raw data matrix for their computation, which makes them unsuitable for use in categorical data. Thus, five internal evaluation criteria, which need only a dissimilarity matrix for their computation, are used. Namely, the Dunn index, the silhouette index, the McClain index, the c-index and the Frey index. Since the Frey index cannot be calculated in every dataset (depending on dataset properties), the remaining four criteria are examined in this paper.

The Dunn index (DU) (Dunn, 1974) assumes that clusters in a dataset are compact and well separated by maximizing the inter-cluster distance while minimizing the intra-cluster distance, see Yang (2012). For the cluster solution with k clusters, it can be expressed by the formula

$$DU(k) = \min_{l \leq g \leq h \leq k} \left(\frac{D(C_g, C_h)}{\max\limits_{l \leq v \leq k} \mathsf{diam}(C_v)} \right),$$

where $D(C_g, C_h)$ is the distance between the g-th and h-th clusters (expressed by a given linkage method), and $\mathsf{diam}(C_v)$ is the maximal distance expressed by a given similarity measure between two objects in the v-th cluster. The Dunn index takes values from zero to infinity. The highest value indicates the optimal cluster solution.

The silhouette index (SI) (Rousseeuw, 1987), also known as the average silhouette width, can be written as

$$SI(k) = \frac{1}{n} \sum_{i=1}^{n} \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

where $a(i)$ is the average dissimilarity of the i-th object to the other objects in the same cluster, and $b(i)$ is the minimal average dissimilarity of the i-th object to other objects in any cluster not containing the i-th object. The silhouette index takes values from −1 to 1. The values close to one indicate well-separated clusters, the values close to minus one suggest badly separated clusters, and values close to zero indicate that the objects in the dataset are often located on the border of two natural clusters. The value zero also indicates single-object clusters.

The McClain index (MC) (McClain and Rao, 1975) is defined as a ratio of the within-cluster and the between-cluster distances

$$MC(k) = \frac{S_w/n_w}{S_b/n_b} = \frac{S_w n_b}{S_b n_w},$$

where n_w is the number of pairs of objects in the same cluster, and n_b is the number of pairs of objects not belonging to the same cluster. S_w is the sum of the within-cluster distances for n_w pairs of objects, and S_b is the sum of the between-cluster distances for n_b pairs of objects. The lowest value of the index indicates the optimal number of clusters.

The c-index (CI) (Hubert and Levin, 1976) is defined as

$$CI = \frac{S_w - S_{\min}}{S_{\max} - S_{\min}}.$$

The statistics n_w and S_w are defined the same way as by the McClain index. $S_{\min}$ and $S_{\max}$ are sums of n_w lowest resp. highest distances across all the pairs of objects. The CI criterion takes values from zero to one, and the optimal number of clusters is attained by its minimum.

3 Experimental Background

This section describes steps that are necessary to set before performing a comparison of the evaluation criteria, namely a process of data generation, a choice of similarity measures, a selection of HCA methods, and the used assessment criteria.

3.1 Data Generation Process

The datasets for the experiment were generated with an aim to cover a wide range of possible situations that can occur. Thus, 81 different dataset settings were used, see Figure 1, which describes the data generation process. The datasets were generated with two to four natural clusters. Three minimal between-cluster distances[2] (0.1, 0.3, 0.5) were used, representing intersecting, partly intersecting and almost non-intersecting clusters. Next, the datasets were generated with three different numbers of variables (4, 7, 10) covering the typical range of clustering of categorical datasets. Based on the empirical experience, three ranges of categories (2–4, 5–7, 8–10) were chosen representing small, medium and

[2]Based on the `sepVal` parameter in the clusterGeneration R package. For instance, `sepVal` = 0.1 represents the low between-cluster distance, where most of the clusters intersect, whereas `sepVal` = 0.5 depicts the high between-cluster distance, where the clusters do not intersect in most of datasets.

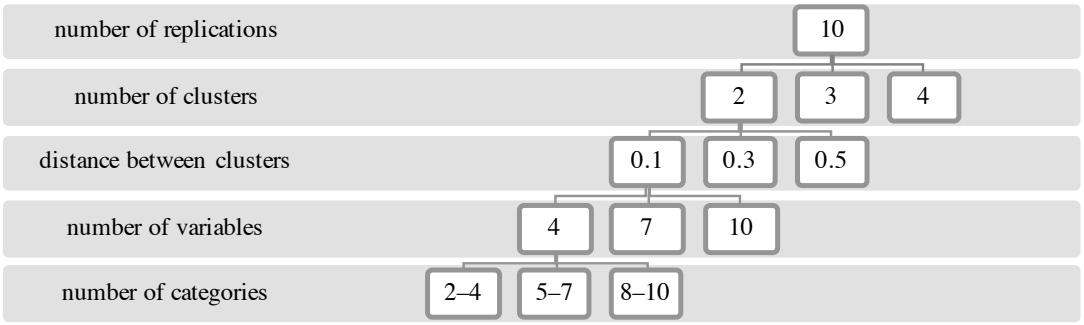

Figure 1: Data generation scheme

large numbers of categories. The numbers of objects in generated datasets were not firmly set; they varied from 300 to 700 cases. To ensure the robustness of the obtained results, each dataset setting combination was replicated ten times. In total, this makes 810 generated datasets that were used for the analysis.

To perform a generation process, an R function with the name `gen_object`, which was developed and described in Šulc (2016), is used. The function depends on the clusterGeneration (Qiu and Joe, 2015) and arules (Hahsler et al., 2017) R packages. The generation is based on a two-step approach. In the first step, a quantitative dataset with multidimensional correlation structure reflecting the given properties (between-cluster distances, the number of clusters, variables and the range of categories) is created. In the second step, the dataset is categorized according to the desired number of categories for each variable in a dataset. The categorization process creates equal-width intervals from the quantitative values of a given variable differing in the numbers of categories. In comparison to an equal-frequency approach, the equal-width approach creates more natural-looking datasets, and moreover, if the categories differ in frequency counts, favorable properties of certain similarity measures can be used.

Figure 2 demonstrates two different dataset generation settings using the `clusplot()` function in the cluster R package (Maechler et al., 2018), which displays the clusters in the two-dimensional space, i.e., there is some loss of data variability. The displayed datasets express 74 % resp. 73.4 % of their original variability. Both the datasets contain three natural clusters, which differ by their minimal between-cluster distance. On the left plot, where `dist` = 0.1 was used, the clusters are largely overlapping, whereas on the right one, they are well separated.

3.2 A Choice of Similarity Measures

Five similarity measures for nominal data, namely SM, ES, IOF, LIN, VE, were chosen for the experiment. The SM (simple matching) measure represents a standardly used approach when determining similarity by datasets characterized by categorical variables. In Šulc (2016) it was found out that the cluster partitions produced by this measure are the same as the partitions created by the majority of similarity measures for binary-coded data based on four frequencies in the 2×2 contingency table, such as the Jaccard coefficient or Sokal and Sneath measures. This finding also corresponds to Todeschini et al. (2012), where 44 similarity measures for binary-coded data were examined, and it was discovered

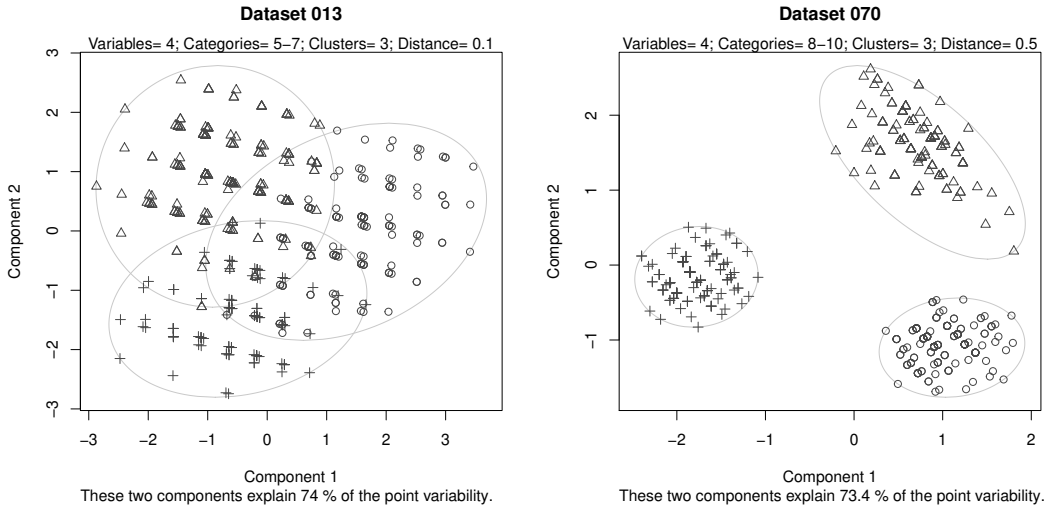

Figure 2: Two generated datasets with different properties

Table 1: Calculation of the used similarity measures for categorical data

Measure	$S_c(x_{ic} = x_{jc})$	$S_c(x_{ic} \neq x_{jc})$	$S(\mathbf{x}_i, \mathbf{x}_j)$	$D(\mathbf{x}_i, \mathbf{x}_j)$
SM	1	0	Eq. (3.1)	Eq. (3.3)
ES	1	$\frac{K_c^2}{K_c^2+2}$	Eq. (3.1)	Eq. (3.4)
IOF	1	$\frac{1}{1+\ln f(x_{ic})\cdot\ln f(x_{jc})}$	Eq. (3.1)	Eq. (3.4)
LIN	$2\ln p(x_{ic})$	$2\ln(p(x_{ic}) + p(x_{jc}))$	Eq. (3.2)	Eq. (3.4)
VE	$-\frac{1}{\ln K_c}\sum_{u=1}^{K_c} p_u \ln p_u$	0	Eq. (3.1)	Eq. (3.3)

that many of them were monotonically dependent and the rest of them were near the monotonical state (the Spearman's Rho ranged from 0.97 to 0.99). Thus, the outputs for the SM measure also represents these similarity measures for binary-coded data.

Next, four similarity measures for nominal data, which provided the best clusters in Šulc (2016), were used. Each of them treats the similarity between two categories differently. The ES measure (Eskin et al., 2002) is based on the number of categories of the c-th variable, whereas the IOF measure (Sparck-Jones, 1972) uses the absolute frequencies of the observed categories x_{ic} and x_{jc}. The LIN measure (Lin, 1998) uses the relative frequencies instead. The VE measure (Šulc, 2016) is based on the variability of the c-th variable expressed by the entropy.

All of the measures can be applied directly to the categorical data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, 2, \ldots, n$ (n is the total number of objects) and $c = 1, 2, \ldots, m$ (m is the total number of variables). The number of categories of the c-th variable is denoted as K_c, absolute frequency as f, and relative frequency as p. Their overview can be found in Table 1, where the column S_c ($x_{ic} = x_{jc}$) presents the similarity computation (or just a value) for matches of categories in the c-th variable for the i-th and j-th objects, and the column S_c ($x_{ic} \neq x_{jc}$) for mismatches of these categories.

At the second level, the total similarity $S(\mathbf{x}_i, \mathbf{x}_j)$ between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$ is determined. For the majority of the examined similarity measures, it is calculated as the

arithmetic mean

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{m}. \tag{3.1}$$

For the LIN measure, the total similarity is expressed as

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m} (\ln p(x_{ic}) + \ln p(x_{jc}))}. \tag{3.2}$$

To compute a proximity matrix, which is required by the majority of software solutions, it is necessary to compute dissimilarities $D(\mathbf{x}_i, \mathbf{x}_j)$ between all pairs of objects, which can be simply obtained from similarities. The dissimilarities are calculated in two ways. For the similarity measures, which take values from zero to one, it is

$$D(\mathbf{x}_i, \mathbf{x}_j) = 1 - S(\mathbf{x}_i, \mathbf{x}_j) \tag{3.3}$$

and for the similarity measures which can exceed the value one, it is

$$D(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{S(\mathbf{x}_i, \mathbf{x}_j)} - 1 \tag{3.4}$$

The $S(\mathbf{x}_i, \mathbf{x}_j)$ and $D(\mathbf{x}_i, \mathbf{x}_j)$ columns in Table 1 show which similarity measures use the particular formulas.

3.3 Methods of Cluster Analysis

To determine the between-cluster distances, three methods of HCA are examined in this paper: the complete linkage, the average linkage, and the single linkage methods. They are commonly used in hierarchical clustering of the categorical data (see, e.g., Lu and Liang, 2008; Morlini and Zani, 2012).

The complete linkage method treats a dissimilarity between two clusters as the dissimilarity between two farthest objects from different clusters. This between-cluster distance usually produces compact clusters with approximately equal diameters. It can be expressed by the formula

$$D(C_g, C_h) = \max_{\mathbf{x}_i \in C_g, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j).$$

The average linkage takes average pairwise dissimilarity between objects in two different clusters. The obtained clusters are often similar to the ones obtained by complete linkage. Its formula can be expressed as

$$D(C_g, C_h) = \frac{1}{n_g n_h} \sum_{\mathbf{x}_i \in C_g} \sum_{\mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j),$$

where n_g and n_h are numbers of objects in the g-th resp. h-th cluster.

The single linkage uses dissimilarity between two closest objects from two different clusters. The formula of this algorithm can be expressed as

$$D(C_g, C_h) = \min_{\mathbf{x}_i \in C_g, \mathbf{x}_j \in C_h} D(\mathbf{x}_i, \mathbf{x}_j).$$

3.4 Evaluation Criteria Assessment

The quality of evaluation criteria will be assessed by the statistic accuracy (AC) and by the mean absolute error (MAE). AC is defined as a percentage of the correctly determined numbers of clusters, expressed as

$$AC = \frac{\sum_{t=1}^{T} I(k_t, K_t)}{n} \cdot 100\%,$$

where k_t is the optimal number of clusters based on given evaluation criteria for the t-th dataset, K_t is the known number of clusters for the t-th dataset, T is the number of datasets, and I is the function which takes the value one in the case of $k_t = K_t$ and the value zero otherwise.

MAE is defined as the average of the absolute differences between optimal numbers of clusters based on a given evaluation criterion and the known numbers of clusters. Low values indicate good stability of an evaluation criterion in its performance and vice versa. MAE is expressed by the formula

$$MAE = \frac{\sum_{t=1}^{T} |k_t - K_t|}{n}.$$

4 Experiment

The experiment consists of two main parts. In the first one, the examined evaluation criteria are assessed regarding their ability to determine the optimal number of clusters. In the second part, the properties of datasets, which significantly associate with the performance of the evaluation criteria, are identified.

The analysis was performed on 810 generated datasets whose generation process was explained in Section 3.1. To each of these datasets, a series of HCAs for two to ten clusters with five examined similarity measures presented in Section 3.2 were applied. The complete, average and single linkage methods of HCA described in Section 3.3 were used. The optimal numbers of clusters based on 11 evaluation criteria presented in Section 2 are then assessed by the AC and MAE statistics. In supplementary material to this paper, a script `run_evaluation.R` containing the whole evaluation process can be found.

4.1 Evaluation of the Optimal Number of Clusters Determination

In this subsection, the optimal number of clusters determination of the 11 examined evaluation criteria will be assessed regarding the used similarity measures and the inherent properties of datasets. The assessment is based on the principle that evenly distributed values of the AC statistic (percentages of the correctly assigned clusters) for a particular evaluation criterion over an examined factor (e.g., similarity measure) indicate that this factor is not associated with the evaluation criterion. Conversely, substantial differences in the AC statistic over the factor values indicate an association between the factor and the criterion. The detailed analysis is limited to results for the average linkage since they provided the best results from all the examined linkage methods. The most important outputs for the complete and single linkages are placed in the Appendix, and they are briefly discussed at the end of this section.

Table 2: AC and MAE statistics broken down by five similarity measures

	AC					MAE				
Crit.	SM	ES	IOF	LIN	VE	SM	ES	IOF	LIN	VE
PSFE	37.8	37.8	37.5	36.9	38.3	1.02	0.98	0.97	1.01	0.93
PSFM	39.1	38.3	39.3	37.3	39.6	1.01	0.98	0.92	1.00	0.91
BK	46.8	40.9	45.3	46.3	44.7	0.74	0.84	0.76	0.73	0.76
BIC1	12.7	38.0	46.0	49.9	44.7	2.95	1.53	1.22	1.04	1.15
BIC2	15.3	16.5	19.9	24.0	19.5	2.37	2.56	2.51	2.18	2.44
DU	29.4	25.3	22.1	23.6	28.0	1.30	1.60	2.12	1.97	1.47
SI	42.4	33.0	41.0	42.3	39.9	1.03	1.59	0.91	1.09	1.17
CI	1.1	4.6	5.6	7.5	0.8	4.80	4.36	4.12	3.48	4.73
MC	33.1	32.5	33.8	32.7	33.1	1.02	1.06	1.05	1.05	1.01

Table 2 shows values of the AC and MAE statistics for the examined internal criteria (Crit.) which were calculated as the averages over all datasets broken down by the five used similarity measures. The criteria AIC1 and AIC2 are not displayed in the output because they provided the same outputs as the more commonly used BIC1 and BIC2 criteria. When looking at the AC values, which express the percentages of correctly determined clusters, it is clear that internal evaluation criteria for categorical data are not nearly as successful as their counterparts for quantitative data, see, e.g., Vendramin et al. (2010), where the accuracy is around 80 %. This is caused by the fact that the clusters in the purely categorical data are much more difficult to recognize since the categorical data have low discriminability compared to quantitative ones. Nevertheless, since nine cluster solutions (two to ten) were investigated, a random guess is 11.1 %, and thus, all the criteria except for CI perform better than that.

The overall best performance among the examined criteria was attained by the BK index, whose AC was around 45 %. The other two variability-based evaluation criteria, PSFM and PSFE, also provided stable but somewhat worse results with AC slightly under 40 %. All three variability-based criteria also have very low MAEs (mostly lower than one); thus, there are very stable in their results. Regarding the likelihood-based evaluation criteria, the BIC1 criterion provided good accuracy (for categorical data) but only by the measures LIN, IOF and VE with MAEs slightly over one. The BIC2 criterion performed poorly by all the similarity measures. This suggests that the entropy-based variants of this type of criteria (BIC1, AIC1) should be preferred. The distance-based evaluation criteria perform rather poorly. The only exception is the silhouette index whose AC is around 40 % (apart from the ES measure).

From Table 2 is also apparent that the MAE values are inversely proportional to the AC scores. The criteria with better accuracy have lower mean absolute errors. The well-performing evaluation criteria have MAEs around one or lower. Overall, the examined evaluation criteria provided the best ACs and MAEs by the LIN measure. Therefore, this similarity measure is going to be used by more detailed analysis and for comparison with the reference SM measure in the rest of the paper.

Table 3 displays the average ACs of the examined evaluation criteria using the LIN

Table 3: ACs of the examined criteria broken by the generated data properties (the LIN measure, average linkage)

	number of clusters			number of variables			range of categories			distance of clusters		
Crit.	2	3	4	4	7	10	2–4	5–7	8–10	0.1	0.3	0.5
PSFE	90.0	13.0	7.8	37.8	37.8	35.2	38.9	35.9	35.9	34.8	37.0	38.9
PSFM	90.0	15.2	6.7	36.7	38.5	36.7	40.0	37.0	34.8	35.2	35.9	40.7
BK	84.8	34.8	19.3	39.3	51.1	48.5	48.1	43.0	47.8	37.8	47.8	53.3
BIC1	81.9	37.8	30.0	37.4	55.6	56.7	47.0	53.7	48.9	36.3	47.0	66.3
BIC2	6.7	35.9	29.3	18.1	26.3	27.4	23.3	23.7	24.8	16.3	20.4	35.2
DU	54.1	7.0	9.6	15.2	28.5	27.0	14.8	28.5	27.4	14.8	23.7	32.2
SI	88.5	20.0	18.5	33.3	47.4	46.3	43.7	44.8	38.5	30.0	40.7	56.3
CI	8.1	10.0	4.4	8.9	6.7	7.0	14.4	3.7	4.4	4.1	9.6	8.9
MC	97.8	0.4	0.0	31.9	33.0	33.3	31.5	33.3	33.3	33.3	33.3	31.5

measure which were averaged over all datasets and broken down by the four different generated dataset properties described in Section 3.1. Considering the first block of the table representing a breakdown based on the number of clusters, it is apparent that all the criteria suffer a significant drop of performance when dealing with more than two natural clusters in a dataset. From the criteria that perform well in determining two clusters (PSFE, PSFM, BK, BIC1, SI, MC), only two criteria (BK, BIC1) exceed 30 % accuracy in three cluster determination, and only the BIC1 criterion reached 30 % accuracy in the four-cluster solution. The second block reveals that some criteria (BK, BIC1, BIC2, DU, SI) substantially improve their performance with an increasing number of variables, the rest of the criteria do not seem to be associated with this factor positively nor negatively. In the third block, the majority of the evaluation criteria is not affected by the number of categories in datasets. The only exceptions are the DU criterion, whose AC is positively associated with the higher numbers of categories, and the CI criterion, which is negatively associated with a higher number of categories. However, both these criteria perform poorly. The last block shows that the majority of the evaluation criteria (BK, BIC1, BIC2, DU, SI, CI) substantially improve their ACs with more distinct clusters in datasets. By the distance 0.5 representing almost non-overlapping clusters, see, e.g., Figure 2, three criteria (BK, BIC1, SI) exceed 50 % accuracy.

Table 4 presents the results of the reference SM measure (and the similarity measures for binary-coded data). Almost all the average accuracies are worse than those by the LIN measure. The largest difference occurs by the BIC1 criterion, which was among the best ones using the LIN measure, whereas it performs very poorly by the SM measure.

4.2 Factors Associated with the Optimal Number of Clusters Determination

As a complement to the analysis in Section 4.1, a series of logistic regression analyses (for each evaluation criterion) was performed. The results of the average linkage with the LIN and SM similarity measures are analyzed. In order to examine dataset properties that are significantly associated with the optimal number of clusters determination, the following logistic regression model was used

$$Y = \frac{e^{\beta_0+\beta_1\cdot clu+\beta_2\cdot var+\beta_3\cdot dist+\beta_4\cdot cat_{2-4}+\beta_5\cdot cat_{5-7}}}{1+e^{\beta_0+\beta_1\cdot clu+\beta_2\cdot var+\beta_3\cdot dist+\beta_4\cdot cat_{2-4}+\beta_5\cdot cat_{5-7}}}.$$

By the dependent variable Y, the values higher or equal to 0.5 stand for correct determination of the optimal number of clusters, and the values lower than 0.5 stand for the unsuccessful determination. As for the independent variables, the model contains three numeric variables, clu (the number of clusters), var (the number of variables), $dist$ (the minimal between-cluster distance), and two dummy variables (cat_{2-4}, cat_{5-7}) based on three ranges of categories (2–4, 5–7, 8–10). The third category is a reference one.

Table 5 presents significant parameters of the logistic regression model that have either positive ($\star$) or negative ($-$) effect on the optimal number of clusters determination of the examined evaluation criteria. In the first block, results for the LIN measure are displayed. The number of clusters proved to be the most significant factor for almost all of the evaluation criteria that is negatively associated with the optimal number of clusters determination (except for the BIC2 criterion). The number of variables proved to

Table 4: ACs of the examined criteria broken by the generated data properties (the SM measure, average linkage)

	number of clusters			number of variables			range of categories			distance of clusters		
Crit.	2	3	4	4	7	10	2–4	5–7	8–10	0.1	0.3	0.5
PSFE	93.3	11.5	8.5	41.9	35.9	35.6	36.3	41.1	35.9	39.6	35.9	37.8
PSFM	93.7	13.3	10.4	41.1	40.7	35.6	39.6	40.7	37.0	38.5	37.8	41.1
BK	91.9	32.6	15.9	48.1	48.1	44.1	44.8	46.7	48.9	43.3	43.7	53.3
BIC1	14.1	11.5	12.6	10.4	11.5	16.3	10.4	14.1	13.7	19.3	10.4	8.5
BIC2	16.7	16.7	12.6	14.4	13.7	17.8	16.3	14.8	14.8	13.3	16.7	15.9
DU	84.8	1.5	1.9	26.7	30.0	31.5	21.5	33.3	33.3	27.4	27.8	33.0
SI	91.9	19.6	15.6	45.2	43.3	38.5	39.6	41.9	45.6	34.1	40.0	53.0
CI	0.7	1.9	0.7	1.1	1.5	0.7	3.3	0.0	0.0	1.1	1.1	1.1
MC	99.3	0.0	0.0	32.6	33.3	33.3	32.6	33.3	33.3	33.3	33.3	32.6

Table 5: Significant parameters of a logistic regression model (LIN and SM measures, average linkage)

	LIN					SM				
Crit.	*clu*	*var*	cat_{2-4}	cat_{5-7}	*dist*	*clu*	*var*	cat_{2-4}	cat_{5-7}	*dist*
PSFE	$---$					$---$				$***$
PSFM	$---$					$---$				$***$
BK	$---$	$**$			$***$	$---$				$***$
BIC1	$---$	$***$			$***$					
BIC2	$***$	$**$			$***$					
DU	$---$	$***$	$---$		$***$	$---$				
SI	$---$	$***$			$***$	$---$				
CI			$***$		$*$					
MC	$---$				$-$					

Note: $*/-$ $p < 0.05$, $**/--$ $p < 0.01$, $***/---$ $p < 0.001$,

be the factor that positively associated with the ability to recognize the optimal number of clusters by the majority of evaluation criteria, namely BK, BIC1, BIC2, DU, and SI. These criteria belonged among the best ones in Section 4.1. The number of categories is significant only by the criteria DU and CI which did not perform well. The increasing between-cluster distance is also a significant factor which increases the success rate of the majority of the evaluation criteria (BK, BIC1, BIC2, DU, SI, CI). By the MC criterion, this effect is negative, but this criterion performed very poorly.

In the second block of Table 5, results for the SM measure are presented. The variability-based measures (PSFE, PSFM, BK) are positively associated with the between-cluster distance. However, their overall performance is not substantially better than by the LIN measure as it was commented in Section 4.1. The majority of the evaluation criteria (PSFE, PSFM, BK, DU, SI) are negatively associated with the number of clusters. It seems that the LIN measure can reflect more properties of datasets than the SM measure can. This is likely the reason why the examined evaluation criteria perform substantially better when this measure was used.

General tendencies of the complete and single linkage methods do not differ substantially from the tendencies presented for the average linkage. However, both the methods are substantially less successful in determining the optimal number of clusters, especially, in datasets with two natural clusters; see Table 6 and Table 7 in the Appendix, where the results for the LIN measure are presented.

5 Conclusion

This paper presented a comparison of 11 internal evaluation criteria for categorical data in hierarchical clustering regarding the optimal number of clusters determination. The evaluation criteria were assessed regarding their accuracy in the optimal number of clusters determination and properties of datasets that significantly associate with the performance of the evaluation criteria. The comparison was performed on 810 generated datasets with

commonly used dataset properties in practice.

The examined internal evaluation criteria performed substantially worse compared to other studies, where the internal evaluation criteria for quantitative data were examined. The accuracy of the best ones was around 45 % while the average accuracy of the quantitative data was around 80 %. One should be especially careful when clustering categorical data that possibly consist of more than two clusters. A solution to this drawback of the examined criteria could be to inspect one lower and one higher number of clusters than the recommended one by some of the recommended criteria. Then, the chance to identify the optimal number of clusters is substantially higher.

In the paper, it was found out that the ability of the evaluation criteria to determine the optimal number of clusters depends on the method of cluster analysis used and the similarity measure used. Based on the results of the performed experiment, the best performance of the evaluation criteria was achieved using the average linkage method. From the five used similarity measures, the evaluation criteria attained the best performance using the LIN measure. Compared with the SM measure, which is commonly used, and its results being the same as using the similarity measures for binary-coded data, the use of the LIN measure increases chances to recognize three or four natural clusters in a dataset. Moreover, with an increasing number of variables or the minimal between cluster distance, it enables to steadily improve the performance of most of the evaluation criteria, which is not the case of the SM measure. Therefore, the use of the LIN measure is recommended.

When examining the association of the generated dataset properties on the optimal number of clusters determination using the logistic regression, it was found out that the increasing number of natural clusters is negatively associated with the accuracy of the majority of the evaluation criteria. On the other hand, an increasing number of variables and the minimal between-cluster distance positively associate with the accuracy of some criteria. The criteria that are positively associated with both the number of variables and the minimal between-cluster distance provide overall better results.

Among the examined evaluation criteria for categorical data clustering, the BIC1 criterion is the most successful in the optimal number of clusters determination if the average linkage with the LIN measure is used. Using complete or single linkage methods, it performs worse than the BK index and similarly as PSFE and PSFM criteria. The BK index is the most robust evaluation criterion from the examined ones. It performs consistently over different methods of hierarchical cluster analysis and similarity measures. Thus, the BIC1 criterion can be used if a researcher chooses the method of hierarchical cluster analysis and a similarity measure. On the other hand, the BK index is best to use if a researcher cannot influence the choice of clustering method and the similarity measure.

Acknowledgement

This work was supported by the University of Economics, Prague under the IGA project No. F4/41/2016.

References

[1] Akaike, H. (1973): Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, 199–213. New York, NY: Springer.

[2] Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M. and Perona, I. (2013): An extensive comparative study of cluster validity indices. *Pattern Recognition*, **46**(1), 243–256.

[3] Bacher, J., Wenzig, K., and Vogler, M. (2004): SPSS TwoStep Cluster – a First Evaluation. Nürnberg: Lehrstuhl für Soziologie.

[4] Bontemps, D. and Toussile, W. (2013): Clustering and variable selection for categorical multivariate data. *Electronic Journal of Statistics*, **7**, 2344–2371.

[5] Biem, A. (2003): A model selection criterion for classification: application to HMM topology optimization. In *Proceeding of Seventh International Conference on Document Analysis and Recognition*, 104–109.

[6] Charrad M., Ghazzali N., Boiteau V., and Niknafs A. (2014): NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, **61**(6), 1–36.

[7] Chen, K. and Liu, L. (2009): "Best K": Critical clustering structures in categorical Datasets. *Knowledge and Information System*, **20**(1), 1–33.

[8] Eskin, E., Arnold, A., Prerau, M., Portnoy, L., and Stolfo, S. V. (2002): A geometric framework for unsupervised anomaly detection. In D. Barbará and S. Jajodia (Eds): *Applications of Data Mining in Computer Security*, 78–100.

[9] Dunn, J. C. (1974): Well separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, **4**(1), 95–104.

[10] de Souto, M. C. P., Coelho, A. L. V., Faceli, K., Sakata, T. C., Bonadia, V., and Costa, I. G. (2012): A comparison of external clustering evaluation indices in the context of imbalanced data sets. In *Proceeding of the 2012 Brazilian Symposium on Neural Networks*, 49–54.

[11] Gini, C. W. (1912): Variability and Mutability. Contribution to the Study of Statistical Distributions and Relations. Studi Economico-Giuridici della R. Universita de Cagliari.

[12] Hahsler, M., Buchta, C., Gruen, B., and Hornik, K. (2017): arules: Mining Association Rules and Frequent Itemsets. R package version 1.5-4. https://CRAN.R-project.org/package=arules

[13] Hubert, L. and Levin, J. R. (1976): A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, **83**(6), 1072–1080.

[14] Light, R.J. and Margolin, B.H. (1971): An analysis of variance for categorical data. *Journal of the American Statistical Association*, **66**(335), 534–544.

[15] Lin, D. (1998): An information-theoretic definition of similarity. In *ICML'98: Proceedings of the 15th International Conference on Machine Learning*, 296–304. San Francisco: Morgan Kaufmann.

[16] Liu, Y, Zhongmou, L., Xiong, H., Gao, X., and Wu, J. (2010): Understanding of internal clustering validation measures. In *Proceedings of 2010 IEEE International Conference on Data Mining*, 911–916.

[17] Löster T. (2013): Modification of CHF and BIC coefficients for evaluation of clustering with mixed type variables. *Research Journal of Economics, Business and ICT*, **8**(2), 9–12.

[18] Lu, Y. and Liang, L. R. (2008): Hierarchical clustering of features on categorical data of biomedical applications. In *Proceedings of the ISCA 21st International Conference on Computer Applications in Industry and Engineering*, 26–31. Hawaii, USA.

[19] Maechler, M., Rousseeuw, P., Struyf, A., Hubert M., and Hornik, K. (2018): cluster: Cluster Analysis Basics and Extensions. R package version 2.0.7-1.

[20] McClain, J. O. and Rao and V. R. (1975): Clustisz: A program to test for the quality of clustering of a set of objects. *Journal of Marketing Research*, **12**, 456–460.

[21] Milligan G. W., and Cooper, M. C. (1985): An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**(2), 159–179.

[22] Morlini I. and Zani S. (2012): A new class of weighted similarity indices using polytomous variables. *Journal of Classification*, **29**(2), 199–226.

[23] Qiu, W. and Joe, H. (2015): clusterGeneration: Random Cluster Generation (with Specified Degree of Separation). R package version 1.3.4. https://CRAN.R-project.org/package=clusterGeneration

[24] Rousseeuw, P. (1987): Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

[25] Řezanková, H., Löster, T., and Húsek, D. (2011): Evaluation of categorical data clustering. In *Advances in Intelligent Web Mastering 3*, 173–182. Berlin: Springer Verlag.

[26] Sparck-Jones, K. (1972): A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, **28**(1), 11–21.

[27] Schwarz, G. (1978): Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

[28] SPSS, Inc. (2001): The SPSS TwoStep Cluster Component. SPSS, Inc.

[29] Šulc, Z. (2016): Similarity Measures for Nominal Data in Hierarchical Clustering. Dissertation thesis, University of Economics, Prague.

[30] Todeschini, R., Consonni, V., Xiang, H., Holliday, J. Buscema, M., and Willett, P. (2012): Similarity coefficients for binary chemoinformatics Data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, **52**(11), 2884–2901.

[31] Vendramin L., Campello, R. J. G. B., and Hruschka, E. R. (2010): Relative clustering validity criteria: A comparative overview. *Statistical Analysis and Data Mining*, **3**(4), 209–235.

[32] Yang, R. (2012): A Hierarchical Clustering and Validity Index for Mixed Data. Dissertation thesis, Iowa State University.

[33] Zhao, Y., Karypis, G., and Fayyad, U. (2002): Evaluation of hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, **10**(2), 141–168.