Cibulková, Jana; Šulc, Zdenek; Sirota, Sergej; Rezanková, Hana

Article

THE EFFECT OF BINARY DATA TRANSFORMATION IN CATEGORICAL DATA CLUSTERING

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Cibulková, Jana; Šulc, Zdenek; Sirota, Sergej; Rezanková, Hana (2019) : THE EFFECT OF BINARY DATA TRANSFORMATION IN CATEGORICAL DATA CLUSTERING, Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, NY, Vol. 20, Iss. 2, pp. 33-47, https://doi.org/10.21307/stattrans-2019-013

STATISTICS IN TRANSITION new series, June 2019
Vol. 20, No. 2, pp. 33–47, DOI 10.21307/stattrans-2019-013

THE EFFECT OF BINARY DATA TRANSFORMATION IN CATEGORICAL DATA CLUSTERING

Jana Cibulková[1], **Zdeněk Šulc**[2], **Sergej Sirota**[3], **Hana Řezanková**[4]

ABSTRACT

This paper focuses on hierarchical clustering of categorical data and compares two approaches which can be used for this task. The first one, an extremely common approach, is to perform a binary transformation of the categorical variables into sets of dummy variables and then use the similarity measures suited for binary data. These similarity measures are well examined, and they occur in both commercial and non-commercial software. However, a binary transformation can possibly cause a loss of information in the data or decrease the speed of the computations. The second approach uses similarity measures developed for the categorical data. But these measures are not so well examined as the binary ones and they are not implemented in commercial software. The comparison of these two approaches is performed on generated data sets with categorical variables and the evaluation is done using both the internal and the external evaluation criteria. The purpose of this paper is to show that the binary transformation is not necessary in the process of clustering categorical data since the second approach leads to at least comparably good clustering results as the first approach.

Key words: hierarchical cluster analysis, nominal variable, binary variable, categorical data, similarity measures, evaluation criteria, generated data.

1. Introduction

The practical importance of cluster analysis increases as the volume of collected data in various fields grows. In the paper, distance-based methods (i.e. methods based on distances or dissimilarities between objects) were chosen for the cluster analysis due to their popularity and ease of implementation in a wide variety of scenarios. Also, according to Charu and Chandan (2013), they can be used with almost any data type, as long as an appropriate measure for given data type exists.

In this paper we focus on hierarchical clustering of objects characterized by categorical variables. This type of data is extremely common in real life. It occurs

[1]Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: jana.cibulkova@vse.cz

[2]Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: zdenek.sulc@vse.cz

[3]Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: sergej.sirota@vse.cz

[4]Department of Statistics and Probability, University of Economics, Prague, Czech Republic. E-mail: hana.rezankova@vse.cz

often in surveys regarding marketing research (important for a market-oriented economy) and in surveys in the field of official statistics (e.g. surveys of living conditions). However, most of the clustering algorithms in the literature focus solely on clustering of numerical data. When clustering nominal data (categorical data that are not numerical nor inherently comparable in any way), a binary transformation is routinely used. This transformation recodes nominal variables into sets of dummy variables and then they are "treated" as if they were binary variables all along. In the process of hierarchical clustering, the distances between objects are expressed based on measures suited for binary data. They are either dissimilarity measures or similarity measures which are transformed into dissimilarities before clustering. Despite the fact that this approach is regarded as a standard procedure when clustering nominal variables, it could cause a loss of information in the data (since it is not one-to-one transformation and it changes underlying distribution of transformed variables) or decrease the speed of the computations (due to dimensionality increase), as demonstrated by Salem et al. (2017).

This transformation, which often creates a data set with substantially larger amount of binary variables, may not be necessary at all, since similarity measures suitable for clustering nominal data exist and can be used instead, see Boriah, Chandola and Kumar (2008), Šulc (2016). These measures are not as well examined as the binary ones and they are usually not implemented in any commercial software and almost never used. In non-commercial software R (R Core Team, 2018), there is a package *nomclust*, that contains several similarity measures suited for clustering nominal data, see Šulc and Řezanková (2015). This package was used for the purpose of clustering categorical data by Ladds et al. (2018).

The main objective of the paper is to determine whether applying binary transformation to categorical data and then using similarity measures for binary data in the process of hierarchical clustering of categorical data (approach one) leads to better-quality clusters than using similarity measures for nominal data (approach two), which can be applied on a data set with categorical variables in its original state. The secondary objective is to evaluate the cluster quality of hierarchical clustering with similarity measures for nominal data compared to the similarity measures for binary data on data sets with purely binary variables. We perform the analysis on 600 generated data sets, where 300 of them are data sets with nominal data and 300 of them are data sets with binary data. The approaches are evaluated using both the internal and the external evaluation criteria. A language and environment for statistical computing R is used for the calculations and the analysis.

2. Similarity measures and linkage method

In this section the chosen similarity (or distance) measures are presented. One group of similarity measures was developed for nominal data and let us refer to those ones as *nominal data measures* in this paper. The other group of similarity measures is suitable for binary data and let us use a term *binary data measures* for them. At the very end of this section, the chosen linkage method is presented.

2.1. Nominal data measures

Seven nominal data measures were used in the experiment:

- ES measure (Eskin et al., 2002),
- IOF measure and OF measure (Sparck-Jones, 1972),
- LIN measure (Lin, 1998),
- LIN1 measure (Boriah et al., 2008),
- VE measure and VM measure (Šulc, 2016),
- SM measure (Sokal and Michener, 1958),
- G1 measure, G2 measure, G3 measure and G4 measure (Boriah et al., 2008).

Let us denote the categorical data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1,\ldots,n$ and $c = 1,\ldots,m$; n is the total number of objects; 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, relative frequency as p, q is a subset of relative frequencies satisfying a set of conditions.

The overview of formulas can be found in Table 1, where the column $S_c\left(x_{ic} = x_{jc}\right)$ presents similarity computation for matches of categories in the c-th variable for the i-th and j-th objects, and the column $S_c\left(x_{ic} \neq x_{jc}\right)$ corresponds to mismatches of these categories. The third column represents the total similarity $S(\mathbf{x}_i, \mathbf{x}_j)$ between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$.

Table 1. Nominal measures overview

Measure	$S_c\left(x_{ic} = x_{jc}\right)$	$S_c\left(x_{ic} \neq x_{jc}\right)$	$S\left(\mathbf{x}_i, \mathbf{x}_j\right)$
ES	1	$\frac{K_c^2}{K_c^2+2}$	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
IOF	1	$\left(1+\ln f(x_{ic}) \ln f(x_{ic})\right)^{-1}$	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
OF	1	$\left(1+\frac{n}{\ln f(x_{ic})}\frac{n}{\ln f(x_{ic})}\right)^{-1}$	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
LIN	$2\ln p(x_{ic})$	$2\ln\left(p(x_{ic})+p(x_{jc})\right)$	$\frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m}[\ln(p(x_{ic})+p(x_{jc}))]}$
LIN1	$\sum_{q\in Q}\ln p(q)$; [5]	$2\ln\sum_{q\in Q} p(q)$;	$\frac{\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})}{\sum_{c=1}^{m}[\ln(p(x_{ic})+p(x_{jc}))]}$
VE	$-\frac{1}{\ln K_c}\sum_{u=1}^{K_c} p_u \ln p_u$	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
VM	$\frac{K_c}{K_c-1}\left[1-\sum_{u=1}^{K_c} p_u^2\right]$	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
SM	1	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
G1	$1-\sum_{q\in Q} p^2(q)$; [6]	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
G2	$1-\sum_{q\in Q} p^2(q)$; [7]	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
G3	$1-p^2(x_{ic})$	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$
G4	$p^2(x_{ic})$	0	$\frac{1}{m}\sum_{c=1}^{m} S_c(x_{ic}, x_{jc})$

[5] $Q \subseteq X_c : \forall q, p(x_{ic}) \leq p(q) \leq p(x_{jc})$
[6] $Q \subseteq X_c : \forall q, p(q) \leq p(x_{ic})$
[7] $Q \subseteq X_c : \forall q, p(q) \geq p(x_{ic})$

In order to compute a proximity matrix, the transformation from similarity into dissimilarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$ is necessary. According to Šulc (2016, pp. 6–10) transformations of similarity measures ES, IOF, OF, LIN, LIN1 (measures which can exceed the value one) to corresponding dissimilarity measures follow the formula:

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

The similarity measures VE, VM, SM, G1, G2, G3, G4 (measures which take values from zero to one) are transformed into corresponding dissimilarity measures using the following formula:

$$D(\mathbf{x}_i, \mathbf{x}_j) = 1 - S(\mathbf{x}_i, \mathbf{x}_j). \tag{2}$$

2.2. Binary data measures

According to Todeschini (2012) binary data measures are often linearly dependent, and thus the majority of them produce the same clusters. Therefore, the binary data measures used in the study are selected in a way that each measure is based on different principle and in some way represents a whole group of (linearly dependent) measures based on given principle. These five binary data measures were chosen for the experiment:

- SMC measure (Sokal and Michener, 1958) is the simple matching coefficient and it is a basic measure used for comparing the similarity and diversity of sample sets,
- EUC measure is the Euclidean distance that is the base for many similarity measures,
- PRS measure (Pearson, 1900) – the Pearson chi-squared statistic is one of many measures based on the Pearson correlation coefficient,
- YUQ measure (Yule, 1912) – Yule's Q represents similarity measures based on odds ratio,
- JAC measure (Jaccard, 1901) – Jaccard similarity measure represents negative match exclusive measures.

Suppose that two objects, $\mathbf{x}_i$ and $\mathbf{x}_j$, are represented by the binary vector form. Let m be the number of variables. There are symbols used for the numbers of variables with certain combinations of categories for objects presented in the Table 2, inspired by Dunn and Everitt (1982). The symbols are used for definitions of binary distance measures in this paper. In Table 2, a is the number of features where the values of $\mathbf{x}_i$ and $\mathbf{x}_j$ are both equal to 1, meaning "positive matches", b is the number of variables where the value of $\mathbf{x}_i$ and $\mathbf{x}_j$ is $(0,1)$, meaning "$\mathbf{x}_i$ absence mismatches", c is the number of variables where the value of $\mathbf{x}_i$ and $\mathbf{x}_j$ is $(1,0)$, meaning "$\mathbf{x}_j$ absence mismatches", and d is the number of variables where both $\mathbf{x}_i$ and $\mathbf{x}_j$ are 0, meaning "negative matches".

Table 2. Symbols used for the numbers of variables with certain combinations of categories for objects $\mathbf{x}_i$ and $\mathbf{x}_j$

$\mathbf{x}_i \backslash \mathbf{x}_j$	1 (Presence)	0 (Absence)
1 (Presence)	a	b
0 (Absence)	c	d

Table 3 provides the overview of formulas of the binary data measures. Some measures were defined as similarity measures, hence the transformation from similarity measure into dissimilarity measure is necessary in order to be able to calculate a proximity matrix. This transformation follows Choi et al. (2010).

Column $S(\mathbf{x}_i, \mathbf{x}_j)$ in the Table 3 represents the total similarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$ if this measure is originally defined as a similarity between objects. $D(\mathbf{x}_i, \mathbf{x}_j)$ in the last column stands for distance between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$.

Table 3. Binary measures overview

Measure	$S(\mathbf{x}_i, \mathbf{x}_j)$	$D(\mathbf{x}_i, \mathbf{x}_j)$
SMC	$\frac{a+d}{a+b+c+d}$	$1 - S(\mathbf{x}_i, \mathbf{x}_j)$
EUC	–	$\sqrt{b+c}$
PRS	$\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$	$\frac{1-S(\mathbf{x}_i, \mathbf{x}_j)}{2}$
YUQ	–	$\frac{2bc}{ad+bc}$
JAC	$\frac{a}{a+b+c}$	$1 - S(\mathbf{x}_i, \mathbf{x}_j)$

2.3. Method of cluster analysis

We applied agglomerative hierarchical cluster analysis (HCA). Its algorithm considers each object to start in its own cluster and at each step the nearest two clusters are combined into a higher-level cluster. This algorithm is usually attributed to Sokal and Michener (1958).

The average linkage method was applied in this analysis since it is a robust method, which is considered a compromise between the single and the complete linkage methods, see (Yim and Ramdeen, 2015). Unlike the single linkage method, the average linkage method is not associated with *chaining phenomenon* and unlike the complete linkage method it is not sensitive to outliers. Also, this method is frequently set as the default one in hierarchical clustering packages. It takes average pairwise dissimilarity between objects in two different clusters. Let us denote $D_{average}(C_k, C_l)$ the distance between cluster C_k and C_l, with the number of objects n_k in the k-th cluster and n_l in the l-th cluster. Then, dissimilarity between two clusters can be expressed by the formula:

$$D_{average}(C_k, C_l) = \frac{\sum_{\mathbf{x}_i \in C_k} \sum_{\mathbf{x}_j \in C_l} D(\mathbf{x}_i, \mathbf{x}_j)}{n_k n_l}. \tag{3}$$

3. Data sets

To achieve the established aims, data sets with nominal and binary variables are generated. In this section, a data generator is introduced and generated data sets are described.

3.1. Generator of nominal data

Data generation is an important part of various research tasks, whether due to lack of real data or, in our case, due to specific requirements given on desired data sets (given number of clusters, variables, variables' categories, ...) that can influence the robustness of the results. Unfortunately, there are not many nominal data generators which can produce data sets with multivariate structure.

In this paper the data generator suitable for the needs of the experiment is used, see (Cibulková and Řezanková, 2018). Each generated data set consists of a given number of clusters, where each cluster corresponds to one sample of a given multivariate distribution. (For the purpose of generating nominal variables, multivariate uniform distribution is desired and multivariate Bernoulli distribution is required in order to generate binary variables.) This idea follows the assumption of finite mixture models from model-based clustering. It is assumed that the population is made up of several distinct clusters, each following a different multivariate probability density distribution, see (Stahl and Sallis, 2012). Hence, the problem of generating data set with given features is reduced to generating samples from given multivariate distributions. To achieve this, NORTA algorithm (Cario and Nelson, 1997) in combination with Cholesky's decomposition (Higham, 2009) is used. Assuming each cluster in the data set is generated from a given multivariate distribution, the generated data set is a mixture of several samples obtained by this approach. This generator allows us to generate numerous data sets with desired features to cover a wide range of data sets "types", making the results of the analysis more robust.

3.2. Data sets with nominal, binary and binarized variables

For the purpose of the analysis, we introduce terms regarding the data sets.

- *Data set with nominal data* is a data set with nominal variables where a number of categories of each variable belongs to the interval $\langle 2, 10 \rangle$. Each column represents one variable.
- *Data set with binary data* is a data set with binary variables, meaning the value of each variable is either 0 or 1. Each column represents one dummy variable.
- *Data set with binarized data* was created by a binary transformation of generated data set with nominal data. Therefore, one variable with K categories from the "original" data set with nominal data transforms into K dummy variables (columns). Hence, this transformation causes that the data set with binarized data contains a lot of zeros and a huge number of columns.

4. Experiment

The experimental part was designed to evaluate two objectives. The first one, connected to the primary aim of the paper, is to determine if better-quality clusters in hierarchical clustering are provided using similarity measures for binary data, which require a binary data transformation, or using similarity measures for nominal data, which can be applied on a data set with nominal data in its original state. The second objective is to evaluate the cluster quality of the similarity measures for nominal data compared to the similarity measures for binary data on data sets with purely binary data. Its outcomes can help to determine if it is meaningful to use nominal data measures on binary data.

4.1. Experiment setting

Using the data generator, which was presented in Section 3, 300 nominal data sets for the main objective and 300 binary data sets for the secondary analysis were generated. A summary of the generated data sets properties is in Table 4.

Table 4. Generated data sets properties

	data sets with nominal data	data sets with binary data
distribution	multivariate uniform distribution	multivariate Bernoulli distribution
number of objects	120–480	120-480
number of categories	2–10	2
number of clusters	4	4
number of variables	10	10
number of replications	300	300

In order to eliminate the influence of the properties which can possibly have effects on the quality of the produced clusters, certain properties were set under control in the performed analyses, while other properties were not set firmly.

The correlation of variables with parameters of multivariate distribution is chosen randomly. The number of objects in a data set varies from 120 to 480 and the number of categories varies randomly from 2 to 10. Data sets with nominal data were generated from multivariate uniform distribution, while multivariate Bernoulli distribution was used for generating data sets with binary data. In both the analyses, the number of clusters was set to four and the number of variables is set to ten to cover typical data set sizes in common clustering tasks. To ensure the robustness of the obtained results, each data set setting combination was replicated 300 times.

4.2. Evaluation criteria

Since the analyses are performed on the generated data sets, and thus objects' cluster memberships are known, the produced clusters can be evaluated using both internal and external evaluation criteria.

For the internal cluster quality evaluation, the variability-based *Pseudo F coefficient based on the mutability* (PSFM) was chosen, see Řezanková et al. (2011).

This coefficient takes into account the within-cluster variability of a data set, which always decreases with the increasing number of clusters. Therefore, the coefficient penalizes an increasing number of clusters. Then, the maximal value indicates the optimal number of clusters. The PSFM criterion can be expressed by the formula

$$PSFM(k) = \frac{(n-k)(WCM(1)-WCM(k))}{(k-1)WCM(k)}, \tag{4}$$

where $WCM(1)$ is the variability in the whole data set with n objects, and $WCM(k)$ the within-cluster variability in the k-cluster solution, which is computed as

$$WCM(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \left(1 - \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g}\right)^2\right),$$

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

The external evaluation of the cluster quality was performed using the *Adjusted Rand Index* (ARI), see Hubert and Arabie (1985), which is commonly used for a comparison of two membership partitions. Compared to the standard Rand index, see Rand (1971), it is corrected for a chance. Similarly to the original measure, which takes values from zero to one, where one indicates that the compared cluster partitions are identical, ARI has a similar range of values, but it can also take small negative values if the *Index* is less than the *Expected index*, see

$$ARI = \frac{\text{Index} - \text{Expected Index}}{\text{Max Index} - \text{Expected Index}} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2} \sum \binom{b_j}{2}]/\binom{n}{2}}, \tag{5}$$

where n_{ij} are the joint frequencies of the contingency table created between two compared partitions, a_i are the row marginal frequencies, and b_j are the column marginal frequencies.

4.3. Evaluation methodology

Values of the two criteria used can be compared not only with their values in different cluster solutions of a certain similarity measure but also with their values in a particular cluster solution for different similarity measures. It can be done by averaging the scores of the evaluation criteria over the examined similarity measures (and/or certain data sets' properties). However, the presented approach can be used only by ARI. The values of PSFM must be processed in a different way since this criterion depends on the number of objects in a data set and also on its initial variability, and thus, it is incomparable in an unadjusted form. Therefore, the proposed procedure uses the two-step rank score approach.

In the first step, clusters produced by HCA with all the examined similarity measures are evaluated. The outcome scores are then ranked in a way that the lowest

rank is assigned to the highest value of the coefficient. Then, the rank scores are averaged in the same way as ARI. The resulting mean rank scores and their standard deviations are considered as the main output which can be displayed in the form of an easily interpretable table. The lower is the mean ranked score of a similarity measure, the better is its clustering performance. The lower the value of the standard deviation, the more stable the clustering performance of a given similarity measure.

4.4. Results of the experiment

Interpretation of the results follows the methodology described in Section 4.3. The measures that have a tendency to create "high-quality" clusters are the ones with low ranks of PSFM index, high values of ARI and low values of PSFM rank's standard deviation.

Table 5 and Table 6 show mean ranks of PSFM, standard deviations of PSFM ranks and average ARIs for each measure. The best values are highlighted in bold writing (the highest values of ARI and the lowest values of PSFM). Table 5 provides a summary of evaluation criteria for data sets with nominal data (these data sets were binarized if the binary data measure was used in the clustering process). Table 6 summarizes the same indices for data sets with binary data (these data sets were generated as data sets with purely binary variables). It is possible to distinguish a type of a measure by the column *Type*, where "B" stands for a binary data measure (black colour) and "N" stands for a nominal data measure (red colour).

Figure 1 and Figure 2 give a visualization of Tables 5 and 6. Axes x and y in the graphs reflect the averages from the tables (average PSFM rank and average ARI) and the size of a grey circle changes according to the standard deviation of PSFM rank. The colours of measures in the figures correspond to the colours in the tables. In these figures, the measures at the bottom right lead to the best clustering solutions, while the measures at the top left lead to the worst clustering solutions.

We can see that in the case of clustering of nominal data and also in the case of clustering binary data, the clustering approach without the binary transformation (using nominal data measures) provides at least as good clustering solutions as the standard approach with binary transformation. According to the chosen evaluation criteria, similarity measures for nominal data and similarity measures for binary data perform comparably well when applied on data sets with nominal (binarized) data. Especially measures EUC, SMC, LIN, VE, VM, SM provided good clustering solutions according to the chosen evaluation criteria. Surprisingly, some similarity measures for nominal data (LIN, LIN1, G3) performed even better than all examined measures for binary data on data sets with binary data. The measure for nominal data LIN steadily leads to the above average clustering solutions when applied to data sets with binary and nominal data. The measures PRS and G4 lead to bellow average clustering solutions. The measures for binary data EUC and SMC handled well high dimensional (binarized) data sets with a lot of zeros. However, they were outperformed by several similarity measures when applied to binary data sets.

Table 5. Experiment results (data sets with nominal/binarized data)

Measure	Type	PSFM		ARI
		Mean	SD	Mean
SMC	B	**7.9**	3.94	**0.566**
EUC	B	**7.9**	3.82	**0.565**
PRS	B	12.9	4.78	0.484
YUQ	B	8.1	4.27	0.548
JAC	B	8.0	3.90	0.567
ES	N	9.7	6.04	0.395
IOF	N	8.2	4.60	0.533
OF	N	8.9	4.59	0.600
LIN	N	**7.9**	4.18	**0.564**
LIN1	N	14.1	4.03	0.512
VE	N	**7.8**	3.94	**0.565**
VM	N	**7.8**	3.91	**0.566**
SM	N	**7.8**	3.77	**0.566**
G1	N	8.8	4.52	0.592
G2	N	8.9	4.49	0.585
G3	N	8.5	4.14	0.580
G4	N	10.0	6.04	0.389

Table 6. Experiment results (data sets with binary data)

Measure	Type	PSFM		ARI
		Mean	SD	Mean
SMC	B	8.1	3.82	0.308
EUC	B	8.5	4.20	0.310
PRS	B	16.9	0.53	0.069
YUQ	B	8.2	4.33	0.303
JAC	B	8.5	4.03	0.308
ES	N	8.1	3.98	0.306
IOF	N	7.9	4.13	0.304
OF	N	8.2	3.88	0.307
LIN	N	**6.5**	4.27	**0.329**
LIN1	N	**6.6**	4.35	**0.329**
VE	N	8.3	3.81	0.307
VM	N	8.2	3.88	0.308
SM	N	8.1	3.84	0.307
G1	N	9.1	5.26	**0.336**
G2	N	11.3	4.70	0.275
G3	N	**6.9**	4.81	**0.336**
G4	N	13.7	4.33	0.200

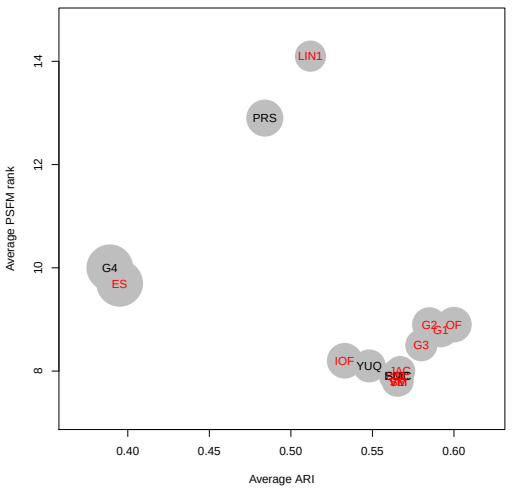

Figure 1: Data sets with nominal (or binarized) data

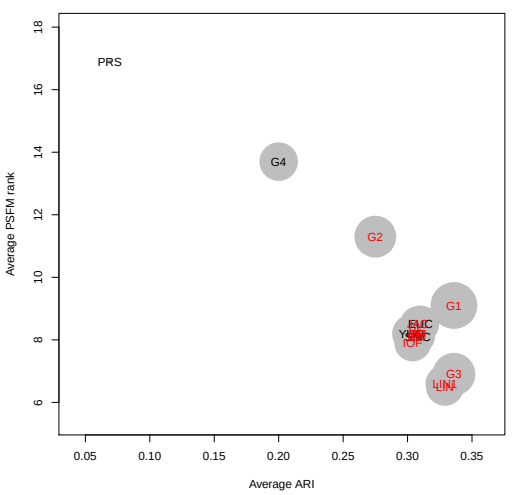

Figure 2: Data sets with binary data

5. Conclusions

In the study we compared two approaches to clustering of categorical data. The first widely used approach performs a binary transformation of the nominal variables into sets of dummy variables and then uses the similarity measures suitable for binary data. The second rarely used approach uses similarity measures developed for the nominal data, hence no data transformation is required. We used internal and external evaluation criteria to determine which of the two approaches creates better quality clusters.

We demonstrated that the binary transformation is not necessary and it is possible to cluster data sets with categorical variables without it. Moreover, according to several internal and external evaluation criteria the approach that uses nominal data measures even leads to "better" clustering results in comparison with clustering solutions obtained by the first approach (clustering data that were transformed by a binary transformation, while using distance measures suitable for binary data) on both types of data sets – data sets with nominal data and data sets with binary data.

Acknowledgements

This work was supported by the University of Economics, Prague under Grant IGA F4/44/2018.

REFERENCES

BORIAH, S., CHANDOLA, V., KUMAR, V., (2008). Similarity measures for categorical data: A comparative evaluation, In Proceedings of the 2008 SIAM International Conference on Data Mining, Society for Industrial, Applied Mathematics, pp. 243–254.

CAIRO, M., NELSON, B., (1997). Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix, Technical Report, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL.

CHARU, C. A., CHANDAN, K. R., (2013). Data Clustering: Algorithms and Applications, Chapman & Hall/CRC.

CHOI, S. S., CHA, S. H., TAPPERT, C. C., (2010). A survey of binary similarity and distance measures,Journal of Systemics, Cybernetics and Informatics, 8 (1), pp. 43–48.

CIBULKOVÁ, J., ŘEZANKOVÁ, H., (2018). Categorical data generator, In International Days of Statistics and Economics 2018. T. Löster and T. Pavelka (eds.) Slaný: Melandrium, Libuše Macáková, pp. 288–296.

DUNN, G., EVERITT, B. S., (1982). An Introduction to Mathematical Taxonomy, Cambridge University Press.

ESKIN, E., ARNOLD, A., PRERAU, M., PORTNOY, L., STOLFO, S. V., (2002). A geometric framework for unsupervised anomaly detection, In Applications of Data Mining in Computer Security, D. Barbará and S. Jajodia (eds.) Boston: Springer, pp. 78–100.

HAHSLER, M., BUCHTA, C., GRUEN, B., HORNIK, K., (2015). Arules: Mining Association Rules and Frequent Itemsets. R package version 1.3-1. https://CRAN.R-project.org/package=arules.

HIGHAM, N. J., (2009). Cholesky factorization, Wiley Interdisciplinary Reviews: Computational Statistics, 1 (2), pp. 251–254.

HUBERT, L., ARABIE, P., (1985). Comparing partitions, Journal of Classification, 2 (1), pp. 193–218.

JACCARD, P., (1901). Étude comparative de la distribuition florale dans une portion des Alpes et des Jura, Bulletin de la Societe Vaudoise des Sciences Naturelles, 37 (142), pp. 547–579.

LADDS, M. A., SIBANDA, N., ARNOLD, R., DUNN, M. R., (2018). Creating functional groups of marine fish from categorical traits, PeerJ 6:e5795.

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

PEARSON, K., (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Philosophical Magazine, Series 5, 50(302), pp. 157–175.

QIU, W., 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.

R CORE TEAM (2018). R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

RAND, W. M., (1971). Objective criteria for the evaluation of clustering methods, Journal of the American Statistical Association, 66 (336), pp. 846–850.

ŘEZANKOVÁ, H., LÖSTER, T., HÚSEK, D., (2011). Evaluation of Categorical Data Clustering, In Advances in Intelligent Web Mastering – 3, Advances in Intelligent and Soft Computing. E. Mugellini, P. S. Szczepaniak, M. C. Pettenati and M. Sokhn (eds.), vol 86. Berlin:Springer, Heidelberg, pp. 173–182.

SALEM, S. B., NAOUALI, S., SALLAMI, M., (2017). Clustering Categorical Data Using the K-Means Algorithm and the Attribute's Relative Frequency, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 11 (6), pp. 708–713.

SOKAL, R., MICHENER, C., (1958). A statistical method for evaluating systematic relationships, University of Kansas Science Bulletin, 38 (2), pp. 1409–1438.

SPARCK-JONES, K., (1972). A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation, 28 (1), pp. 11–21.

STAHL, D., SALLIS, H., (2012). Model-based cluster analysis, In Wiley Interdisciplinary Reviews: Computational Statistics, 4 (4), pp. 341–358.

ŠULC, Z., (2016). Similarity measures for nominal data in hierarchical clustering. Dissertation thesis, Prague: University of Economics.

ŠULC, Z., ŘEZANKOVÁ, H., (2015). Nomclust: An R package for hierarchical clustering of objects characterized by nominal variables, In International Days of Statistics and Economics 2018. T. Löster and T. Pavelka (eds.) Slaný: Melandrium, pp. 1581–1590.

TODESCHINI, R., CONSONNI, V., XIANG, H., HOLLIDAY, J., BUSCEMA, M., 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), pp. 2884–2901.

YULE, G U., (1912). On the methods of measuring association between two attributes, Journal of the Royal Statistical Society, 49 (6), pp. 579–652.

YIM, O., RAMDEEN, K. T., (2015). Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data, The Quantitative Methods for Psychology, 11 (1), pp. 8–21.