STATISTICS IN TRANSITION new series, Autumn 2014
Vol. 15, No. 4, pp. 591–610

EVALUATION OF SELECTED APPROACHES TO CLUSTERING CATEGORICAL VARIABLES

Zdeněk Šulc[1], **Hana Řezanková**[2]

ABSTRACT

This paper focuses on recently proposed similarity measures and their performance in categorical variable clustering. It compares clustering results using three recently developed similarity measures (IOF, OF and Lin measures) with results obtained using two association measures for nominal variables (Cramér's V and the uncertainty coefficient) and with the simple matching coefficient (the overlap measure). To eliminate the influence of a particular linkage method on the structure of final clusters, three linkage methods are examined (complete, single, average). The created groups (clusters) of variables can be considered as the basis for dimensionality reduction, e.g. by choosing one of the variables from a given group as a representative for the whole group. The quality of resulting clusters is evaluated by the within-cluster variability, expressed by the WCM coefficient, and by dendrogram analysis. The examined similarity measures are compared and evaluated using two real data sets from a social survey.

Key words: variable clustering, nominal variables, association measures, similarity measures.

1. Introduction

When dealing with high dimensional data, reduction of the number of variables is often desired. It can spare both the computational time and costs for gathering the information in the future. The use of principal component analysis or factor analysis, as described, for example, by Jolliffe (2002), or their categorical counterparts, such as correspondence analysis Greenacre (2010), is very popular. These methods provide additional information about a data set, variables of which have significant loadings on a shared vector, see Palla et al. (2012). An approach based on multiple correspondence analysis for large data sets

[1] Department of Statistics and Probability, University of Economics, Prague. W. Churchill sq.4, 130 67 Praha 3, Czech Republic. E-mail: zdenek.sulc@vse.cz.
[2] Department of Statistics and Probability, University of Economics, Prague. W. Churchill sq.4, 130 67 Praha 3, Czech Republic. E-mail: hana.rezankova@vse.cz.

is presented by D'Enza and Greenacre (2012). Another way to achieve the dimensionality reduction of a data set can be to create groups of similar variables using cluster analysis. One variable of each group can be chosen as a representative for further analysis. Hierarchical cluster analysis represents the basic approach used for variable clustering, see Gordon (1999), Gan et al. (2007). It is based on a proximity matrix, which contains dissimilarities of analyzed variables taken pairwise. More sophisticated approaches are represented, for example, by model-based clustering, see Chavent et al. (2010); Everitt et al. (2011). In R software, one might find a few variable clustering procedures in a package named ClustOfVar, see Chavent et al. (2012). The practical use of variable clustering can be found in various fields of use, e.g. in questionnaires surveys, actuarial sciences, chemistry, gene expression analysis, see Palla et al. (2012), or in getting rid of redundant variables in predictive models, see Payne and Edwards (1999).

The paper focuses on comparison of two kinds of similarity measures which can be used in variable clustering with binary or nominal variables. The first ones are the association measures, Cramér's V and the uncertainty coefficient, which express the dependency between two variables based on the chi-square statistic and the ANOVA method. The second kind is represented by recently developed similarity measures, IOF, OF and Lin, which were originally proposed for object clustering, but have been adjusted for variable clustering in this paper. Clustering with both kinds of measures is going to be compared with the simple matching coefficient, which is commonly used in categorical data clustering and thus it can serve as a reference measure.

The IOF, OF and Lin measures have never been evaluated for variable clustering; they have only been studied for object clustering so far. Moreover, the evaluations of these measures were performed only with the known cluster membership, see Boriah et al. (2008), Chandola et. al. (2009); thus cluster analysis was treated more like a classification problem with supervised learning. Moreover, both publications were focused on the outlier detection performance of the similarity measures.

In this paper, two data sets from a social survey are analyzed. The quality of clusters, obtained using different similarity measures, is evaluated from aspects of both the within-cluster variability, measured by the WCM (within-cluster mutability) coefficient, and the dendrogram analysis. To minimize the influence of clustering algorithm on clustering performance of the similarity measures, clusters obtained by three linkage methods are compared and evaluated.

The rest of the paper is organized as follows. Section 2 introduces the association and other similarity measures. Section 3 describes evaluation criteria of cluster quality. The application of theoretical approach to real data is presented in Section 4. The final results are summarized in the Conclusion.

2. Nominal variable clustering

A basic approach to variable clustering is to create a dissimilarity matrix, which contains dissimilarities of analyzed variables taken pairwise, and then to apply agglomerative hierarchical cluster analysis. A dissimilarity measure can be derived from a similarity measure. Many similarity measures have been proposed for categorical data. One can use association measures for nominal variables, see Anderberg (1973), or similarity measures determined for objects characterized by nominal variables. There are also several other approaches, for example, in Chavent et al. (2010), where the adjustment of existing centre-based method for categorical variable clustering is presented. It is not possible to compare all approaches or all measures; therefore, we focus only on the selected ones.

Three linkage methods of hierarchical clustering are applied in this paper: *complete method* (CLM), *single method* (SLM) and *average method* (ALM). In CLM, the dissimilarity between the furthest variables from two different clusters is considered as the distance between these clusters. SLM takes the dissimilarity between the nearest variables from two different clusters for this purpose, and ALM takes the average distance of all dissimilarities between variables from two different clusters.

2.1. Association measures

Different types of association measures for nominal variables are used in multivariate analysis. Some of them are based on Pearson's chi-squared statistic, some on the principle of dependence measurement in the ANOVA method.

The measures based on the chi-square statistic compare observed and expected counts under the hypothesis of independence; these counts are frequencies of combinations of categories of two nominal variables. Pearson's coefficient of contingency, Cramér's V and the phi coefficient belong to this group. In this paper, *Cramér's V* is applied because it takes values from the interval [0, 1] and takes into account the numbers of categories. It is calculated according to the formula

$$V = \sqrt{\chi^2 / n(q-1)}\,, \tag{1}$$

where χ^2 is Pearson's chi-squared statistic, n is the number of surveyed objects and q is a minimum number of categories of two analyzed variables. If at least one variable is dichotomous, then values of Cramér's V equal the values of the phi coefficient. Cramér's V can be transformed into a dissimilarity measure by subtracting its value from 1.

In the ANOVA method, a directional dependence is considered. In such a case, a symmetric measure is calculated as the harmonic mean of two asymmetric measures. There are two symmetric coefficients for nominal variables derived from asymmetric measures which are based on the principle of ANOVA: the

lambda coefficient and the uncertainty coefficient. The former one is based only on frequencies of modal categories, the latter one takes into account frequencies of all combinations of categories. Therefore, *the uncertainty coefficient* is applied in our experiments. It takes values from the interval [0, 1] and it is based on the entropy as a variability measure. For the *c*-th and *d*-th variables it is calculated as

$$U_{cd} = \frac{2 \cdot (H_c + H_d - H_{cd})}{H_c + H_d}, \tag{2}$$

where H_c (H_d) is the entropy of the *c*-th (*d*-th) variable and H_{cd} is the within-group entropy. Generally, the entropy H is expressed as

$$H = -\sum_{u=1}^{h} p_u \ln p_u , \tag{3}$$

where p_u is a relative frequency of the *u*-th category and h is the number of categories if for all u $p_u \neq 0$. In the case of $p_u = 0$, the corresponding addend equals 0 for this u. The uncertainty coefficient can be transformed into a dissimilarity measure by subtracting its value from 1.

More association measures for variable clustering can be found in Řezanková (2014).

2.2. Recently developed similarity measures

Compared with association measures, which are based on frequencies in a contingency table, the other similarity measures considered in this paper compare categories taken pairwise for each object individually. The term *the other similarity measures* covers the recently developed similarity measures (IOF, OF and Lin) and the overlap measure, which serves as a reference measure. All these measures have a drawback which is that all analyzed variables must have the same number of categories and the categories must have the same meaning. The reason is as follows: if categories across the variables did not have the same meaning, it would make no sense to compare them. For this reason the same number of categories is considered.

All formulas in this paper are based on the data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, 2, ..., n$ and $c = 1, 2, ..., m$ (n is the total number of objects, m is the total number of variables).

Originally, the *IOF* (inverse occurrence frequency) measure comes from an information retrieval, where it used to serve to determine a relative number of documents containing a specific word, see Sparck-Jones (1972, 2002). The original measure was designed to deal only with binary variables; later, it was adjusted to deal with nominal variables as well. The measure was constructed to assign higher weights to mismatches on less frequent values and lower weights to mismatches on more frequent values. When determining similarity between variables $\mathbf{x}_c$ and $\mathbf{x}_d$ for the *i*-th object, it can be expressed as

$$S_i(x_{ic}, x_{id}) = \begin{cases} 1 \text{ if } x_{ic} = x_{id} \\ \dfrac{1}{1 + \ln f(x_{ic}) \cdot \ln f(x_{id})} \text{ otherwise} \end{cases}, \tag{4}$$

where $f(x_{ic})$ is a frequency of the category x_{ic} of the i-th object. Dissimilarity between variables $\mathbf{x}_c$ and $\mathbf{x}_d$ is expressed as

$$D(\mathbf{x}_c, \mathbf{x}_d) = \frac{1}{\dfrac{\sum_{i=1}^{n} S_i(x_{ic}, x_{id})}{n}} - 1. \tag{5}$$

The *OF* (occurrence frequency) measure has an opposite system of weights to the IOF measure. It assigns higher weights to mismatches on more frequent values and otherwise, i.e.

$$S_i(x_{ic}, x_{id}) = \begin{cases} 1 \text{ if } x_{ic} = x_{id} \\ \dfrac{1}{1 + \ln \dfrac{m}{f(x_{ic})} \cdot \ln \dfrac{m}{f(x_{id})}} \text{ otherwise} \end{cases}. \tag{6}$$

Dissimilarity can be determined using Equation (5).

The *Lin* measure, which was introduced by Lin (1998), represents an information-theoretic definition of similarity based on relative frequencies. It was derived from theoretic assumptions about similarity. The emphasis was put on the universality of use; thus, it can be used in various situations including determination of similarity between ordinal values. It assigns higher weights to more frequent categories in the case of a match and lower weights to less frequent categories in the case of a mismatch, i.e.

$$S_i(x_{ic}, x_{id}) = \begin{cases} 2 \cdot \ln p(x_{ic}) \text{ if } x_{ic} = x_{id} \\ 2 \cdot \ln(p(x_{ic}) + p(x_{id})) \text{ otherwise} \end{cases}, \tag{7}$$

where $p(x_{ic})$ expresses a relative frequency of the category x_{ic} of the i-th object. The dissimilarity measure is defined as

$$D(\mathbf{x}_c, \mathbf{x}_d) = \frac{1}{\dfrac{\sum_{i=1}^{n} S_i(x_{ic}, x_{id})}{\sum_{i=1}^{n} (\ln p(x_{ic}) + \ln p(x_{id}))}} - 1 \cdot \tag{8}$$

Clustering with the measures mentioned above is compared with results obtained using the *overlap* measure, which takes into account only whether two observations match or not. When determining similarity between variables $\mathbf{x}_c$ and $\mathbf{x}_d$ for the i-th object, it assigns value 1 if the variables match and value 0 otherwise.

$$S_i(x_{ic}, x_{id}) = \begin{cases} 1 \text{ if } x_{ic} = x_{id} \\ 0 \text{ otherwise} \end{cases}. \tag{9}$$

The dissimilarity measure is defined as

$$D(\mathbf{x}_c, \mathbf{x}_d) = 1 - \frac{\sum_{i=1}^{n} S_i(x_{ic}, x_{id})}{n}. \tag{10}$$

Unlike recently developed similarity measures, the overlap measure does not take into account frequency distribution of categories of a given object, which could serve as an important factor for determining similarity between variables. The comparison of the above mentioned coefficients applied for an object clustering with respect to the within-cluster variability is described in Šulc and Řezanková (2014).

3. Evaluation criteria of final clusters

In this paper, the quality of final clusters is evaluated from the aspects of the WCM (within-cluster mutability) coefficient and by the dendrogram analysis.

The within-cluster variability is an important indicator of cluster quality. With an increasing number of clusters, the within-cluster variability decreases, so the clusters become more homogenous. In this paper, the measurement of the within-cluster variability is based on the *Gini coefficient*, which determines the variability (mutability) of nominal variables. It is expressed by the following equation

$$G_{gi} = 1 - \sum_{u=1}^{h} \left(\frac{n_{giu}}{m_g} \right)^2, \tag{11}$$

where m_g is the number of variables in the g-th cluster (g = 1, ..., k), n_{giu} is the number of variables in the g-th cluster by the i-th object with the u-th category (u = 1, 2, ..., h; h is the number of categories). After standardization of this coefficient with the aim to get values from 0 to 1, and its extending for n objects and k clusters, it can be expressed in a form of the *normalized within-cluster mutability coefficient*:

$$WCM(k) = \frac{1}{n} \frac{h}{h-1} \sum_{g=1}^{k} \frac{m_g}{m} \sum_{i=1}^{n} \left(1 - \sum_{u=1}^{h} \left(\frac{n_{giu}}{m_g} \right)^2 \right), \tag{12}$$

where m is the number of variables. The *WCM* coefficient is based on the G' measure, which was proposed by Řezanková et al. (2011) for the purpose of evaluation of object clustering.

When clustering a relatively small number of variables, the dendrogram analysis can be very helpful. Dendrograms visualize the process of agglomerative

hierarchical clustering calculation. They have a form of charts, which have the examined variables, e.g. on the Y axis, and the distance between clusters on the X axis. They can be cut at any point to get a particular cluster solution.

4. Real data application

To illustrate the influence of selected association and other similarity measures on variable clustering, two variable sets, which come from the research *Men and Women with a University Degree*, are chosen. This survey was conducted by the *Institute of Sociology of the Academy of Sciences of the Czech Republic*, see the archives of the institute (http://archiv.soc.cas.cz).

The following software was used for the analysis: Matlab, IBM SPSS Statistics, STATISTICA and MS Excel. In Matlab, proximity matrices for all similarity measures were computed. In IBM SPSS, hierarchical cluster analyses using CLM, SLM and ALM were performed. In STATISTICA, dendrograms were created. In MS Excel, evaluation criteria for cluster quality evaluation were computed.

4.1. Description of the variable sets

Two batteries of questions were chosen for the analysis. The first battery consists of 9 variables; all with two possible answers *yes* or *no*. The questions are: *From family reasons, have you ever:* p27a – *worked part-time*, p27b – *worked in shifts*, p27c – *worked flextime*, p27d – *changed a job*, p27e – *changed a profession*, p27f – *moved*, p27g – *refused a job offer*, p27h – *refused a promotion offer*, p27i – *cheated at work?* The cases with missing values were omitted, so answers from 1,904 respondents were included.

The second battery deals with gender equality. It contains 9 variables, which all have three possible answers: *women have better opportunities than men*, *men and women have approximately equal opportunities* and *men have better opportunities than women*. The variables are the following: p13a – *to get a job*, p13b – *to have better salary for the same job*, p13c – *to get a leadership*, p13d – *to be a director*, p13e – *to be promoted*, p13f – *for a salary increase*, p13g – *to gain benefits*, p13h – *to have authority*, p13i – *to keep a job*. There is one additional variable with the name: p12 – *a chance of success* which has the same categories as the previous battery of questions. For this reason, it can be added to the set of variables. Overall, answers from 1,886 respondents were used.

4.2. Binary variable clustering

Table 1 presents values of the WCM coefficient for the solutions with two to five clusters for CLM, computed for the set of questions with binary answers. The quality of a particular cluster solution can be evaluated according to the within-cluster variability expressed by the WCM coefficient. The lower the value of WCM, the better the cluster solution. For the two-cluster solution, most of the

measures, except for the Lin measure, provide the same results, i.e. 0.366. For cluster solutions for three and more clusters, the best results are provided by the recently developed similarity measures, i.e. IOF, Lin and OF, which have the same results. They are followed by the overlap measure and further by the both association measures.

Table 1. Values of the WCM coefficient for clustering of binary variables (CLM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.366	0.320	0.255	0.186
Coefficient U	0.366	0.320	0.254	0.186
IOF measure	0.366	0.297	0.232	0.168
OF measure	0.366	0.297	0.232	0.168
Lin measure	0.375	0.297	0.232	0.168
Overlap measure	0.366	0.301	0.236	0.172

Another approach to evaluate the clustering performance is to use dendrograms, which are presented in Figure 1. When looking at the dendrograms, it is apparent that they can be separated into three groups from the point of view of the clustering structure. The first one comprises both the association measures, the second one includes the recently developed similarity measures and the last one contains only the overlap measure. Similarity measures in a particular group provide similar results. Since data dimension reduction is the primary goal of variable clustering, low-cluster solutions are preferred.

When using SLM, as shown in Table 2, one might see that the results are very different from the results achieved by CLM. Generally, they are all worse. There are apparent interesting changes in behaviour of the similarity measures. Both association measures perform better than the recently developed similarity measures from the point of view of their within-cluster variability and the interpretation of dendrograms. Moreover, using SLM, the advantage of recently developed similarity measures, which is based on taking into account frequency distribution of categories, is not apparent in the results. Thus, their results are very similar to the overlap measure, which is also demonstrated by the similar structure of dendrograms of clustering with these measures in Figure 2. The best clusters are provided by Cramér's V in the three-cluster solution.

Table 2. Values of the WCM coefficient for clustering of binary variables (SLM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.378	0.299	0.232	0.186
Coefficient U	0.372	0.333	0.232	0.186
IOF measure	0.376	0.307	0.245	0.172
OF measure	0.376	0.307	0.245	0.190
Lin measure	0.376	0.307	0.245	0.190
Overlap measure	0.376	0.307	0.245	0.190

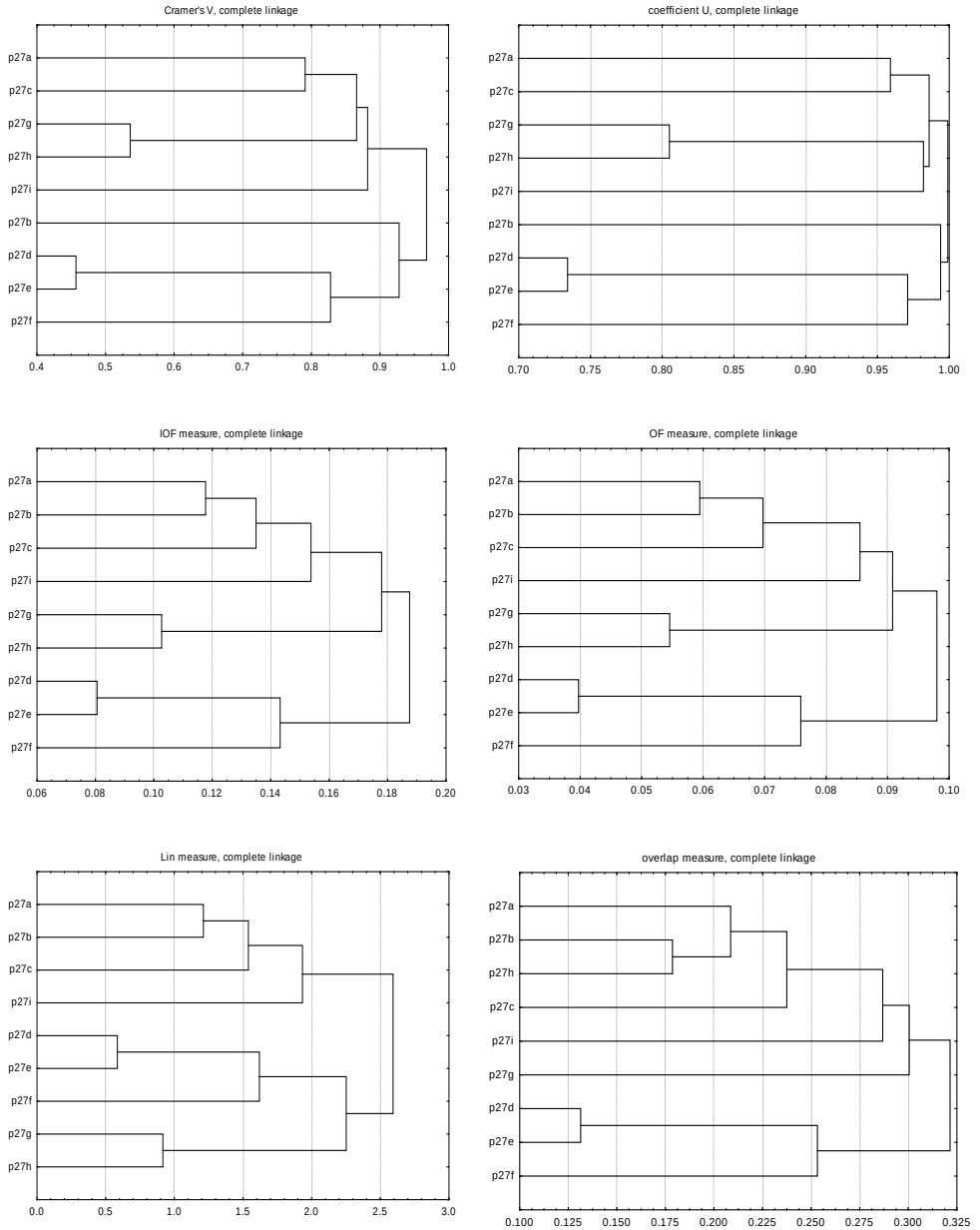

Figure 1. Dendrograms for clustering of binary variables (CLM)

It is important to note that the distances between pairs of variables are differentiated much worse by SLM than by CLM. This fact can cause a bad assignment of clusters into new ones when performing the agglomerative process,

because there are very small differences in their distance by SLM. Especially, such situations are noticeable by the uncertainty coefficient and the IOF measure in Figure 2.

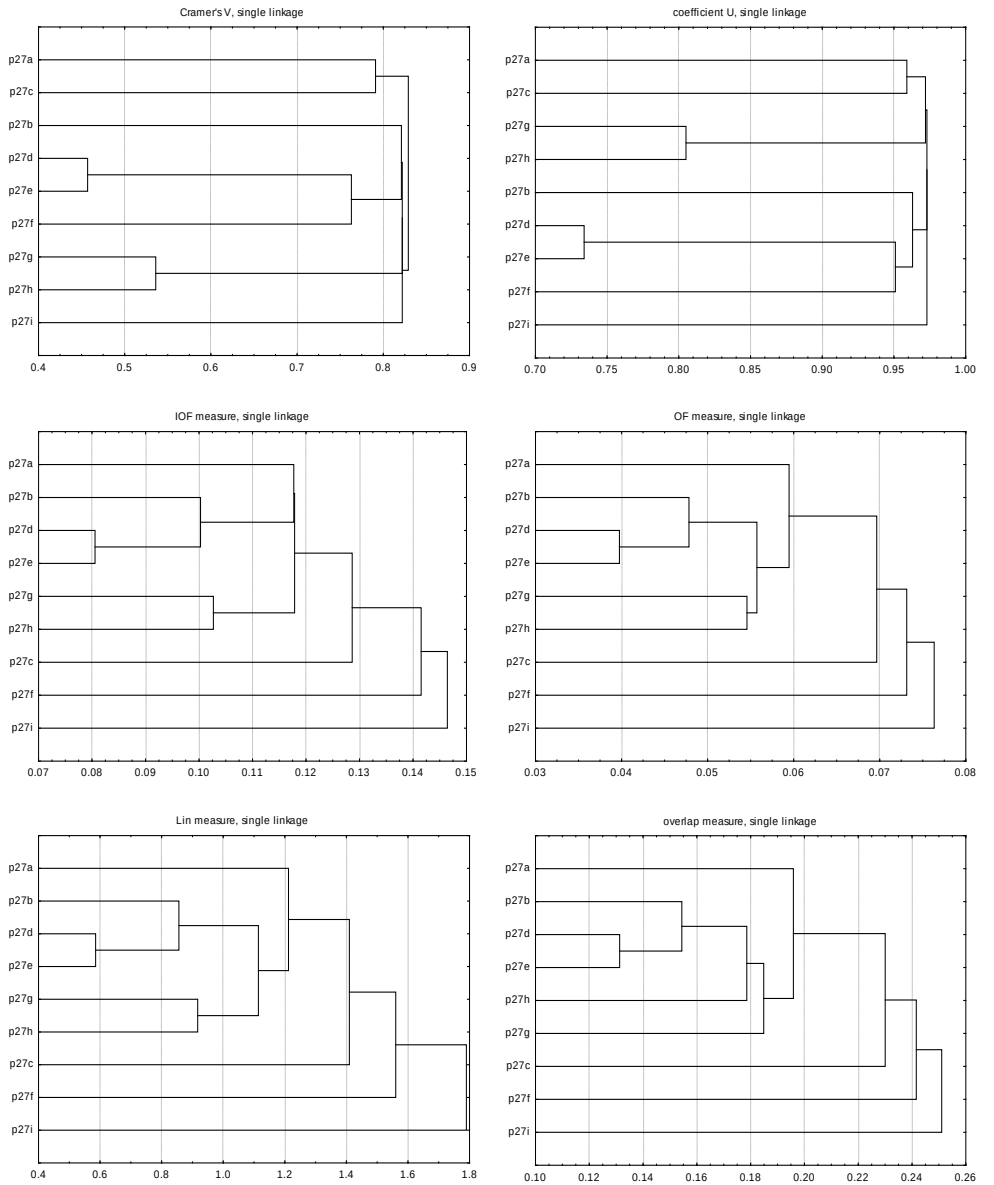

Figure 2. Dendrograms for clustering of binary variables (SLM)

When evaluating the WCM results obtained on the basis of ALM, one can observe that they lie somewhere in between the results of CLM and SLM, as shown in Table 3. However, their structure is much more similar to the one of

CLM, as demonstrated in Figure 3. When examining the dendrograms, one can notice that the distances between clusters are not as large as by CLM, but they are considerably larger than by SLM. The best clusters are provided by the IOF measure. Actually, they are exactly the same as when using CLM.

Table 3. Values of the WCM coefficient for clustering of binary variables (ALM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.366	0.320	0.254	0.186
Coefficient U	0.366	0.333	0.232	0.186
IOF measure	0.366	0.297	0.232	0.168
OF measure	0.375	0.307	0.234	0.172
Lin measure	0.366	0.300	0.232	0.168
Overlap measure	0.375	0.307	0.238	0.177

In the binary variable set, the best clusters are provided by IOF only when using SLM Cramér's V provides better results. Unfortunately, it is not that they are good but because the other measures perform much worse. All the recently developed similarity measures have satisfying results when using CLM or ALM. In the end, the three-cluster solution of the IOF measure by CLM was chosen. The clusters look as follows. In the first cluster, there are variables regarding the kind of work (p27a – *worked part-time*, p27b – *worked in shifts*, p27c – *worked flextime*, p27i – *cheated at work*). The second cluster summarizes variables concerning changing a job (p27d – *changed a job*, p27e – *changed a profession*, p27f – *moved*). The third cluster describes variables regarding a refusal of a good offer in a job (p27g – *refused a job offer*, p27h – *refused a promotion offer*).

4.3. Three-category variable clustering

The within-cluster variability for two- to five-cluster solutions using CLM for three-category variables is contained in Table 4. The results are not as unambiguous as by the binary variables. In the two-cluster solution, the best results provide both the OF and the overlap measure. In the three-cluster solution, there is a different situation; both IOF and Lin have the best results. All the association measures provide worse results in comparison to other similarity measures, which have very similar results of the WCM coefficient.

Table 4. Values of the WCM coefficient for clustering of three-category variables (CLM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.416	0.354	0.287	0.208
Coefficient U	0.427	0.352	0.287	0.208
IOF measure	0.385	0.317	0.259	0.208
OF measure	0.381	0.322	0.261	0.196
Lin measure	0.385	0.317	0.259	0.194
Overlap measure	0.381	0.321	0.260	0.195

Looking at the dendrograms in Figure 4, it is apparent that they can be divided into three groups according to the clustering structure. The first group contains both the association measures, Cramér's V and the uncertainty coefficient. These measures have a tendency to create unbalanced clusters; all of them provide at least one cluster comprising only one variable. The second group includes IOF and Lin, and in the last group, there are OF and overlap. According to dendrograms interpretation, the best results are provided by the Lin measure, which has, except for the five-cluster solution, the same results as the IOF measure.

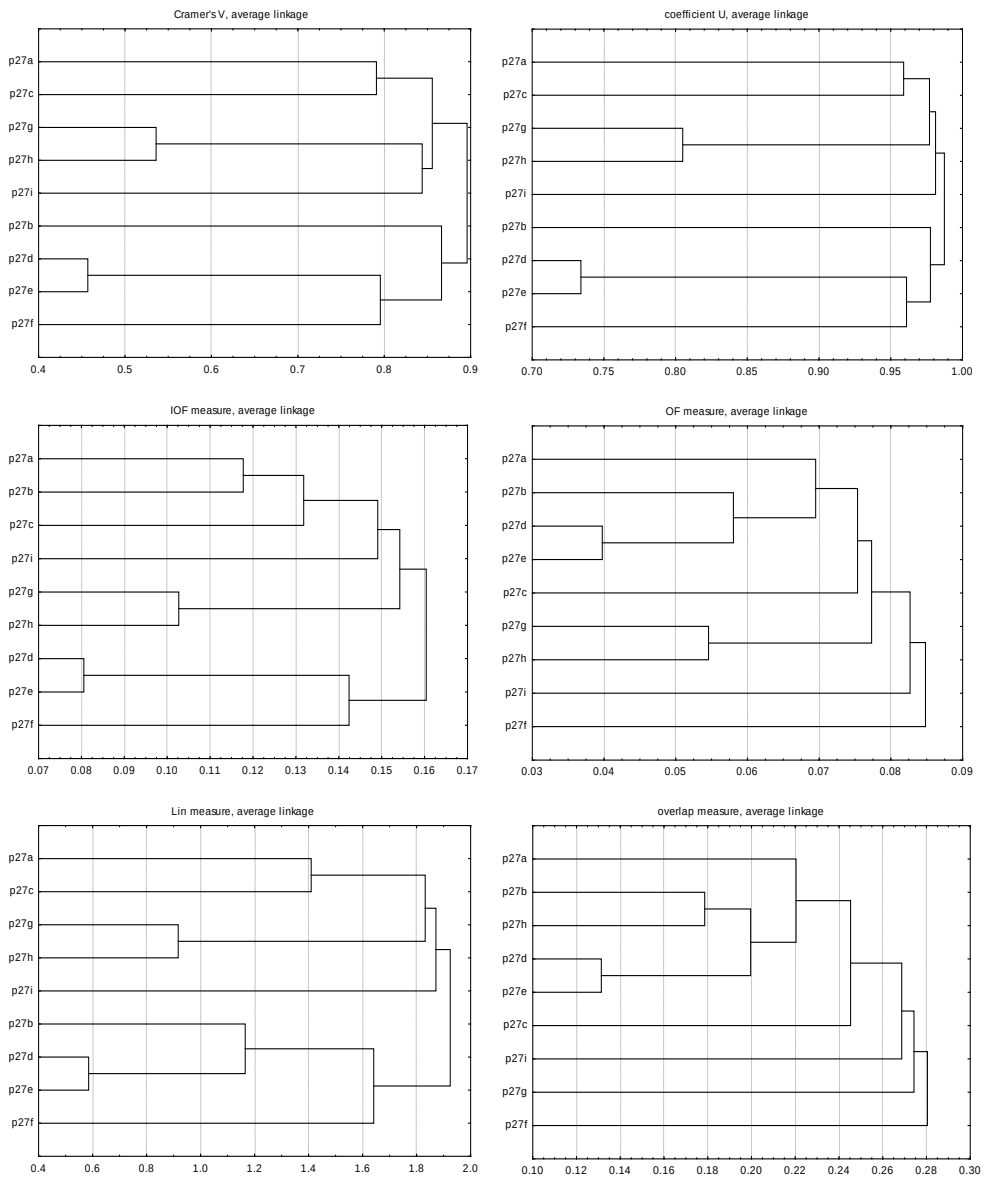

Figure 3. Dendrograms for clustering of binary variables (ALM)

When using SLM, the within-cluster variability of similarity measures in a particular cluster solution is expressed in Table 5. Similarly as by the binary data set, the clustering results are much worse than by CLM. Except for the IOF measure, all other similarity measures provide very unbalanced clusters, which often contain only one variable.

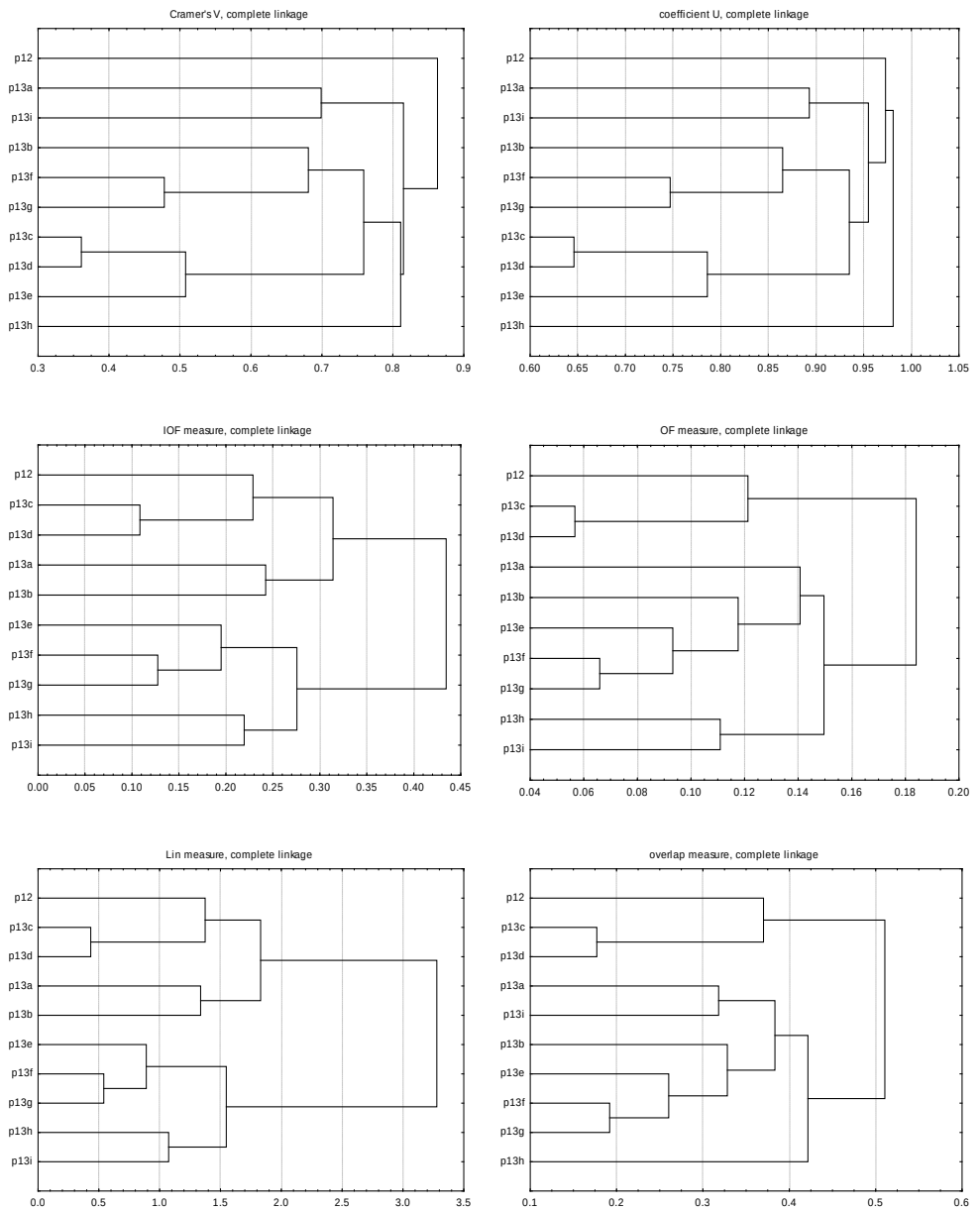

Figure 4. Dendrograms for clustering of three-category variables (CLM)

Table 5. Values of the WCM coefficient for clustering of three-category variables (SLM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.416	0.352	0.295	0.248
Coefficient U	0.427	0.352	0.287	0.240
IOF measure	0.381	0.329	0.267	0.202
OF measure	0.416	0.358	0.295	0.198
Lin measure	0.416	0.358	0.257	0.202
Overlap measure	0.416	0.358	0.295	0.198

According to the dendrograms in Figure 5, the OF and overlap measures provide clusters in a similar way. Again, the advantage of recently developed similarity measures, which take into account the frequency distribution of categories, does not seem to have a big importance by SLM. The best clusters are provided by the IOF measure, but they do not reach the quality of the same measure by CLM.

The values of the WCM coefficient for ALM are displayed in Table 6. They are very similar to those provided by CLM; they differ only in details. The overlap measure has the best results across all cluster solutions. It is closely followed by the recently developed similarity measures and then by the association measures.

Table 6. Values of the WCM coefficient for clustering of three-category variables (ALM)

	WCM(2)	WCM(3)	WCM(4)	WCM(5)
Cramér's V	0.416	0.352	0.295	0.209
Coefficient U	0.427	0.352	0.287	0.208
IOF measure	0.381	0.317	0.267	0.208
OF measure	0.381	0.322	0.263	0.198
Lin measure	0.385	0.323	0.259	0.202
Overlap measure	0.381	0.316	0.256	0.198

When looking at the dendrograms displaying the ALM clustering in Figure 6, one might see that some of them have a similar structure to CLM (the uncertainty coefficient, the overlap measure, and all the recently developed similarity measures). Thus, some measures provide similar results to SLM and some to CLM. The best results are provided by the overlap measure.

Generally, in the three-category variable set, the best results are provided by the IOF measure. Outputs based on this measure are not the best in all cluster solutions; however, they are very robust in most situations. Actually, the best results by CLM, the Lin measure, and by ALM, the overlap measure, were the same as those provided by the IOF measure. By SLM, the IOF measure performed beyond competition.

The two-cluster solution obtained by CLM with the IOF measure was considered to be the best one. The first cluster deals with variables concerning getting a job (p13a – *to get a job*, p13b – *to have better salary for the same job*, p13c – *to get a leadership*, p13d – *to be a director* and p12 – *a chance of success*). The second cluster consists of variables regarding getting a better position in a respondent's job: (p13e – *to promote*, p13f – *for a salary increase*, p13g – *to gain benefits*, p13h – *to have authority*, p13i – *to keep a job*).

5. Conclusion

In this paper, clustering performance of two kinds of similarity measures was examined: the association measures for nominal variables and the other similarity measures originally proposed for objects characterized by nominal variables. There were two main aspects of the comparison. Firstly, the final cluster solutions were evaluated from the point of view of the within-cluster variability; secondly, on the basis of dendrograms and judgments of the researcher. For the analysis, sets of binary and three-category variables were chosen. The influence of different types of linkage methods on resulting clusters was also examined.

Overall, six similarity measures were evaluated in this paper. There were two association measures and four other similarity measures. The association measures, Crammer's V and the uncertainty coefficient, focus on general dependence between two variables when determining their similarity. However, this way of similarity measuring may lead to a loss of some part of information, and thus, to worse dissimilarity determination. The results of the within-cluster mutability (WCM) coefficient and clusters unbalanced by this measures confirmed such a scenario. Therefore, the use of association measures is not suitable for clustering of nominal variables in cases where other possibilities can be considered.

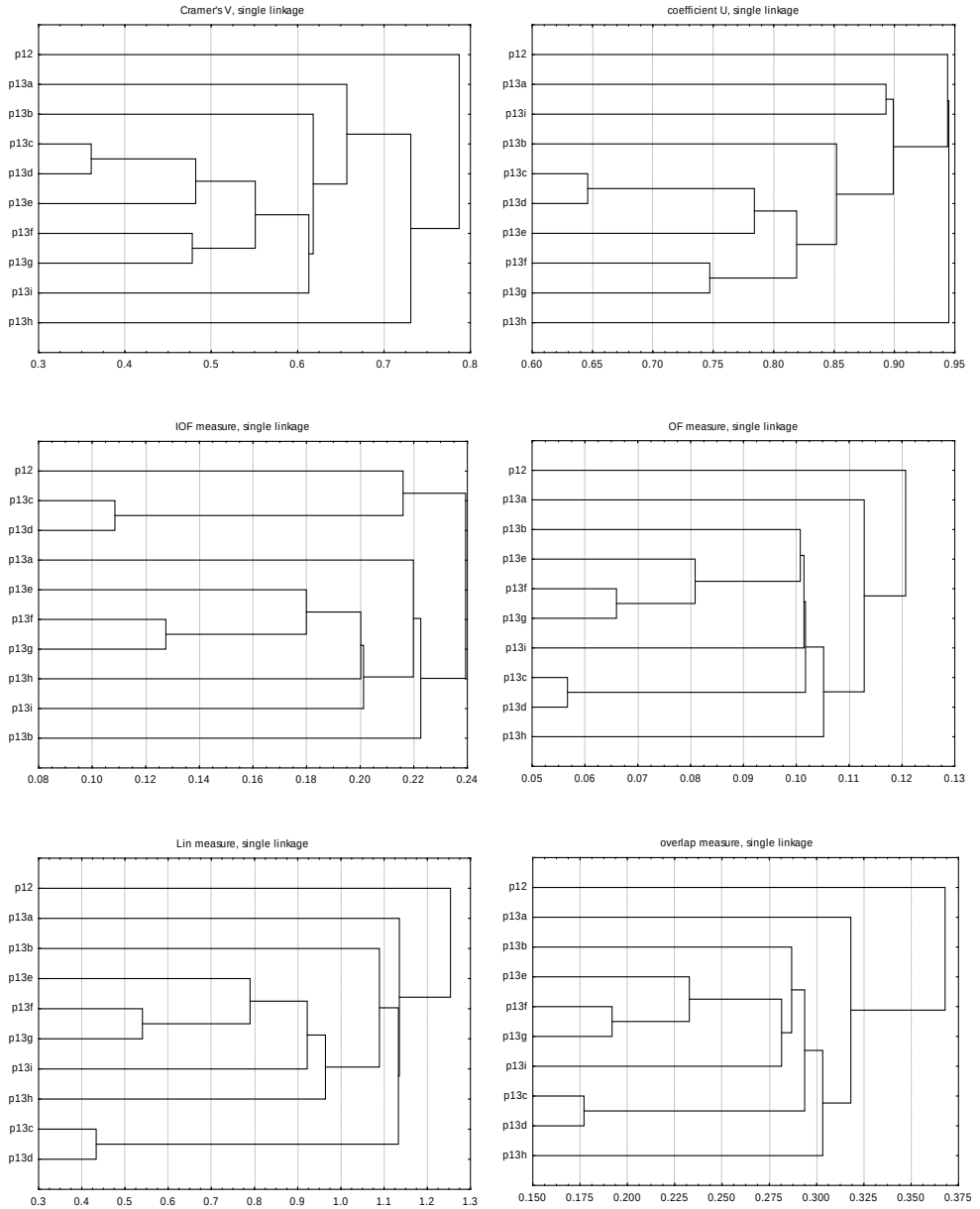

Figure 5. Dendrograms for clustering of three-category variables (SLM)

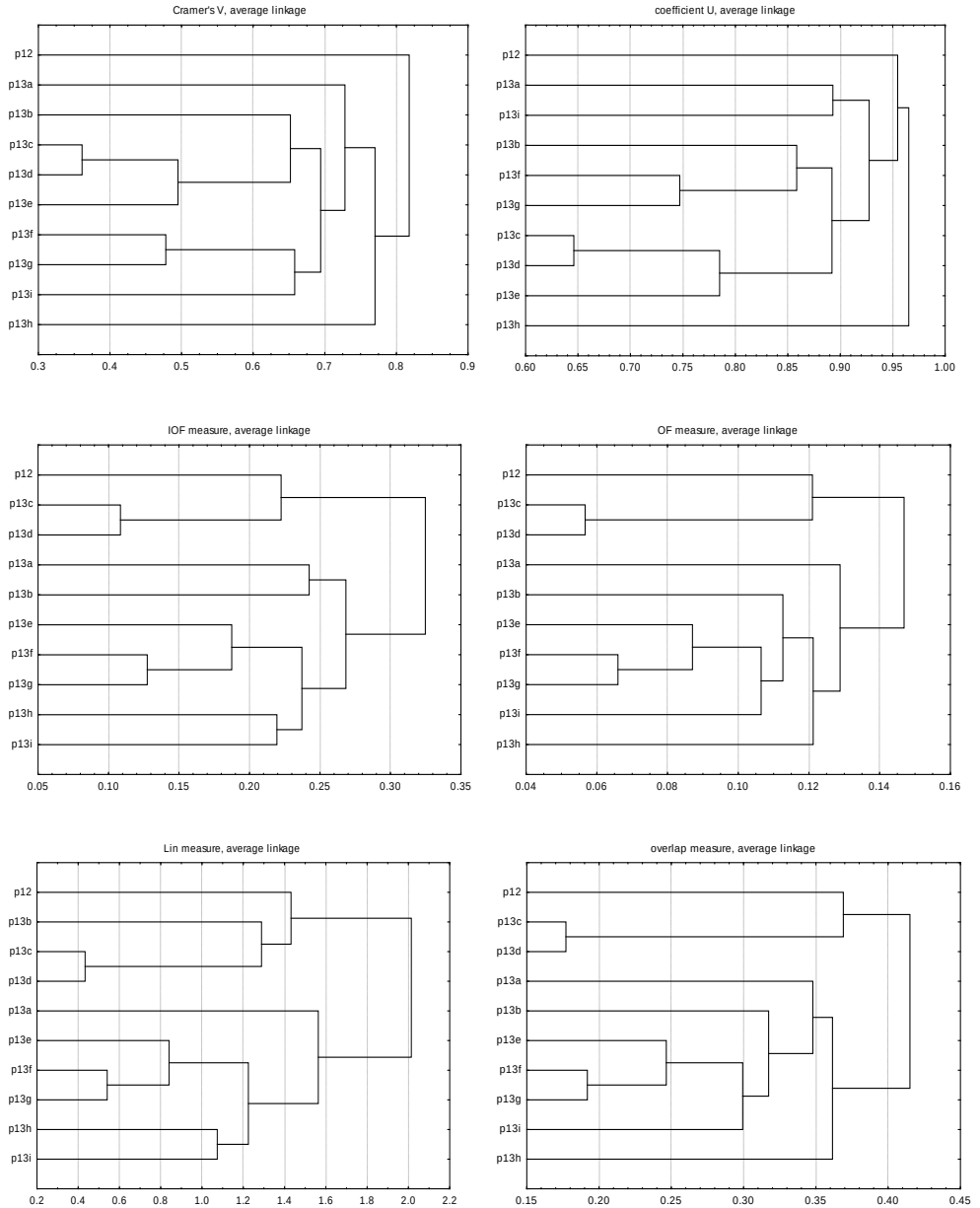

Figure 6. Dendrograms for clustering of three-category variables (ALM)

Four other similarity measures were examined: IOF, OF, Lin and overlap, which compare categories taken pairwise for each object individually, and which all require the same number of categories with the same meaning. Those measures differ mainly in their weight systems. The IOF measure assigns higher weights to

mismatches on less frequent categories which allows it to be more sensitive to outliers in a data set. This approach proved to be more successful in comparison with the OF measure, which uses exactly the opposite weight system, which puts lower weights to those outliers. The Lin measure, as well as the OF measure, assigns lower weights to less frequent categories in the case of a mismatch, but more than that, it also assigns higher weights to more frequent categories in the case of a match. This makes its results very robust in comparison with the OF measure. The overlap measure offers no weight system. This measure provided similar results of the WCM coefficient with the rest of other similarity measures; however, the crucial difference was in cluster quality of resulting clusters. They were unbalanced and their dendrogram interpretation was worse than the rest of the other measures. On the whole, the IOF and Lin measures provided very good clusters of variables in both data sets from the aspects of the WCM coefficient as well as the dendrogram interpretation. Therefore, the use of one of these measures is highly recommended for variable clustering.

When comparing the three linkage methods, the best results are provided by the complete one. It provides good differentiation of clusters; thus, it is easy to cut a dendrogram at a given point. Further, it creates clusters of a similar size, which is in accordance with reduction of a data set. The single linkage method provides very different results in comparison to the complete and average linkage methods. Moreover, the adjustments of recently developed similarity measures, which take into account frequency distribution of categories, do not seem to have any strong influence because of this method. On the whole, this method offers the worst results of all the examined linkage methods; therefore, it cannot be recommended for variable clustering. Thus, the complete or average linkage method should be preferred.

Acknowledgement

This work was supported by the University of Economics, Prague under the project IGS F4/104/2014.

REFERENCES

ANDERBERG, M. R., (1973). Cluster Analysis for Applications. Academic Press, New York.

BORIAH, S., CHANDOLA, V., KUMAR, V., (2008). Similarity measures for categorical data: a comparative evaluation. In: Proceedings of the 8th International Conference on Data Mining. SIAM, pp. 243–254.

CHANDOLA, V., BORIAH, S., KUMAR, V., (2009). A framework for exploring categorical data. In: Proceedings of the 9th International Conference on Data Mining. SIAM, pp. 187–198.

CHAVENT, M., KUENTZ, V., LIQUET, B., SARACCO, L., (2012). ClustOfVar: An R package for the clustering of variables. Journal of Statistical Software, 50(13):1–16. Available at: <http://arxiv.org/abs/1112.0295> [Accessed: 16 October 2014].

CHAVENT, M., KUENTZ, V., SARACCO, J., (2010). A partitioning method for the CLUSTERING of categorical variables. In: Locarek-Junge, H., Weihs, C., eds, Classification as a Tool for Research. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin Heidelberg, pp. 91–99.

D'ENZA, A. I., GREENACRE, M. J., (2012). Multiple correspondence analysis for the quantification and visualization of large categorical data sets. In: Advanced Statistical Methods for the Analysis of Large Data-Sets. Springer, Berlin Heidelberg, pp. 453–463.

EVERITT, B. S., LANDAU, S., LEESE, M., STAHL, D., (2011). Cluster Analysis, 5th edn, Wiley, Chichester.

GAN, G., MA, C., WU, J., (2007). Data Clustering: Theory, Algorithms, and Applications, ASA-SIAM, Philadelphia.

GORDON, A. D., (1999). Classification, 2nd edn, Chapman & Hall/CRC, Boca Raton.

GREENACRE, M. J., (2010). Correspondence analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2(5):613–619.

JOLLIFFE, I. T., (2002). Principal Component Analysis, 2nd edn, Springer, New York.

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

PALLA, K., KNOWLES, D. A., GHAHRAMANI, Z., (2012). A nonparametric variable clustering model. In: Pereira, F., Burges, C. J. C., Bottou, L., Weinberger, K. Q., eds, Advances in Neural Information Processing Systems 25. NIPS Foundation. Available at: <http://papers.nips.cc/paper/4579-a-nonparametric-variable-clustering-model.pdf> [Accessed 16 October 2014].

PAYNE, T. R., EDWARDS, P., (1999). Dimensionality reduction through correspondence analysis. Available at: <http://eprints.soton.ac.uk/263091/> [Accessed 16 October 2014].

ŘEZANKOVÁ, H., LÖSTER, T., HÚSEK, D., (2011). Evaluation of categorical data clustering. In: Mugellini, E., Szczepaniak, P. S., Pettenati, M. C. et al., eds, Advances in Intelligent Web Mastering 3. Springer Verlag, Berlin, pp. 173–182.

ŘEZANKOVÁ, H., (2014). Nominal variable clustering and its evaluation. In: Proceedings of the 8th International Days of Statistics and Economics. Melandrium, Slaný, pp. 1293–1302. Available at: < http://msed.vse.cz/msed_2014/article/276-Rezankova-Hana-paper.pdf > [Accessed 5 November 2014].

SPARCK-JONES, K., (1972, 2002). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28(1):11–21. Later: Journal of Documentation, 60(5):493–502.

ŠULC, Z., ŘEZANKOVÁ, H., (2014). Evaluation of recent similarity measures for categorical data. In: Proceedings of the 17th International Conference Applications of Mathematics and Statistics in Economics. Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, Wroclaw, pp. 249–258. Available at: < http://www.amse.ue.wroc.pl/papers/Sulc,Rezankova.pdf> [Accessed 5 November 2014].