# NOMINAL VARIABLE CLUSTERING AND ITS EVALUATION

Hana Řezanková

**Abstract**

The paper evaluates clustering of nominal variables using different similarity measures. The created clusters can serve for dimensionality reduction by choosing one of the variables from a group of similar variables as a representative for the whole group. A suitable way for variable clustering is to apply hierarchical cluster analysis based on a proximity matrix expressing relationships for all pairs of variables. For measurement of similarity of nominal variables special measures have to be used, e.g. contingency coefficients, symmetric variants of directional dependence measures, measures of agreement, and coefficients determined for measurement of a similarity of objects characterized by nominal variables (for variables with categories of the same meaning). In the paper, the following measures are applied: Cramer's V, the symmetric uncertainty coefficient, the kappa coefficient of agreement, and the simple matching (overlap) coefficient. In addition, the Jaccard coefficient is used for binary variables and the agreement is measured by the Hamann coefficient for this type of variables. The complete method of hierarchical cluster analysis is applied to sets of variables from a sociological research. For evaluation of the created clusters, the within-cluster variability based on the Gini coefficient is considered.

**Key words:** cluster analysis, variable clustering, nominal variables, similarity measures, evaluation of clustering

**JEL Code:** C19, C38, C52

## Introduction

In some researches, mainly those, which are based on questionnaire surveys, nominal variables are often analyzed. These variables cannot be ordered and their analyses differ from analyses of ordinal and quantitative variables, see e.g. (Řezanková and Löster, 2013). In the paper, cluster analysis of such variables is discussed. This multivariate method is useful for identification of groups of similar variables based on the answers of respondents. The created groups can be considered as a basis for dimensionality reduction, e.g. by choosing one of the variables from a given group as a representative for the whole group. This approach applied for ordinal variables has been published e.g. by Prokop and Řezanková (2011). It is an alternative approach to other methods for dimensionality reduction, see e.g. (Li et al. 1995; Maaten et al. 2008; Bohdalová and Greguš, 2011; Frolov et al., 2014).

The basic term in cluster analysis is *similarity*. It can be expressed by means of measures of similarity, dissimilarity or distance. For measurement of similarity in nominal variable clustering, contingency coefficients can be applied, see e.g. (Anderberg, 1973). Further, association measures which are symmetric variants of asymmetric coefficients can be used. For variables with categories of the same meaning (the same number of categories is supposed), the coefficient of agreement and also the simple matching (overlap) coefficient can be applied. The proximity matrix based on some of the coefficients mentioned above serves as an input for hierarchical cluster analysis.

The aim of this paper is to evaluate nominal variable clustering using different similarity measures. The within-cluster mutability is applied for this purpose. It is based on the Gini coefficient, which is determined for measurement of variability of nominal variables.

# 1 Methodology

Agglomerative hierarchical clustering is applied for the analyses presented in this paper. In this type of cluster analysis, objects and clusters of objects are joined step by step into one cluster. The distance between two furthest objects from two different clusters is considered as the distance between these clusters (this method is called *the complete linkage*). Some other approaches for clustering of categorical variables are presented in (Chavent et al., 2010).

For the analysis, the IBM SPSS Statistics and STATISTICA systems were used. The evaluation coefficients were calculated in the first mentioned system in a simple way based on standard commands.

## 1.1 Similarity measures for nominal variables

Similarity of two nominal variables can be expressed e.g. by measures of dependence. If the variables have the same number of categories with the same meaning, then the measure of agreement or the simple matching coefficient can be applied. We can mention examples of these variables from the living condition surveys: indicators if a household has certain durables (a washing machine, a color TV, a telephone, a personal car, a computer etc.) having three categories: yes, no – cannot afford, no – other reason.

Several *measures of dependence* are applied in categorical data analysis. One group of them is based on the Pearson chi-square statistic, which compares observed frequencies of categories of two variables and expected counts under the hypothesis of independence. There are the Pearson coefficient of contingency, Cramer's V and the phi coefficient in this group. *Cramer's V* has values from the interval ⟨0; 1⟩. It is expressed as

$$V = \sqrt{\frac{\chi_P^2}{n(q-1)}}, \qquad (1)$$

where $\chi_P^2$ is the Pearson chi-square statistic, $n$ is a number of studied objects and $q$ is a minimum number of categories of two analyzed variables. If at least one variable is dichotomous, then the values of Cramer's V equal the values of the phi coefficient. The maximum value of the Pearson contingency coefficient depends on numbers of categories of variables and it is less than 1. For a computation of a proximity matrix, a dissimilarity measure is needed. Cramer's V has values from 0 to 1, therefore it can be transformed into a dissimilarity measure by subtracting the value of the coefficient from 1. This measure is used for the further analysis. Similarly, a dissimilarity measure of all coefficients mentioned below is obtained by subtracting the absolute value of the coefficient from 1.

The second group of similarity measures is based on the principle of a dependence measurement in the ANOVA method when the between-groups variability is compared to the total variability (the ratio of two values expressing the variability is calculated). This principle is determined for a directional dependence. The measures have values from the interval ⟨0; 1⟩. In some cases, symmetric measures are calculated as the harmonic average of two asymmetric measures. The *uncertainty coefficient* based on the entropy as a variability measure is used for experiments presented in this paper. For the *j*th and *l*th variables this coefficient is denoted $U_{jl}$ and it is calculated as

$$U_{jl} = \frac{2 \cdot (H_j + H_l - H_{jl})}{H_j + H_l}, \qquad (2)$$

where $H_j$ ($H_l$) is the entropy of the *j*th (*l*th) variable and $H_{jl}$ is the within-group entropy (the mean entropy within rows or columns of a contingency table).

The other coefficients based on the ANOVA method are not suitable because the tau coefficient exists only in the variants of two asymmetric measures (any symmetric variant does not exist) and the lambda coefficient is based only on modal categories and it does not take a total frequency distribution into account.

Agreement of categories of two variables can be expressed by the *kappa coefficient*, which is based on expected counts under the hypothesis of independence. It is supposed that the total agreement of two variables occurs when the values of these variables are the same for each object. The kappa coefficient is expressed as

$$\kappa = \frac{\sum_{u=1}^{K} n_{uu} - \sum_{u=1}^{K} e_{uu}}{n - \sum_{u=1}^{K} e_{uu}}, \qquad (3)$$

where $n_{uu}$ are observed diagonal frequencies of a contingency table ($u = 1, 2, \ldots, K$, $K$ is a number of categories) and $e_{uu}$ are frequencies expected under the hypothesis of independence. This coefficient has values from the interval $\langle -1; 1\rangle$. The value 1 means the total agreement and 0 means that the observed diagonal frequencies equal the frequencies expected in case variables are independent.

The basic measure proposed for the evaluation of similarity of objects characterized by nominal variables is the *simple matching coefficient,* which is also called the *overlap measure*. If all variables from a certain group have the same number of categories and the categories have the same meaning, then this measure can be applied for evaluation of variable similarity. Let us denote the similarity of variables $X_j$ and $X_l$ as $s_{jl}$. For calculation of the overlap measure, the values in the *j*th and *l*th columns of input matrix **X** are compared for all studied objects. Evaluation of relationships of the values for the *i*th object (the *i*th row of matrix **X**) is denoted as $s_{ijl}$. If $x_{ij} = x_{il}$, then $s_{ijl} = 1$, otherwise $s_{ijl} = 0$. The overlap measure $O_{jl}$ is calculated as the arithmetic mean, i.e.

$$O_{jl} = \frac{\sum_{i=1}^{n} s_{ijl}}{n}. \qquad (4)$$

This measure has values from the interval $\langle 0; 1\rangle$.

In case of two dichotomous variables, many other coefficients of similarity or dissimilarity can be used. Let us denote the observed frequencies for a combination of categories $n_{uv}$, where $u$ = 1, 2 and $v$ = 1, 2. Then the *overlap measure* can be expressed as

$$O = \frac{n_{11} + n_{22}}{n}. \qquad (5)$$

*Cramer's V* is calculated according to the formula

$$V = \frac{|n_{11}n_{22} - n_{12}n_{21}|}{\sqrt{(n_{11} + n_{12}) \cdot (n_{11} + n_{21}) \cdot (n_{12} + n_{22}) \cdot (n_{21} + n_{22})}}. \qquad (6)$$

The value of this measure equals to the absolute values of correlation coefficients (the values of all correlation coefficient are the same in case of dichotomous variables).

For the evaluation of agreement of categories, the *Hamann coefficient* is applied for dichotomous variables. It is expressed as

$$H = \frac{(n_{11} + n_{22}) - (n_{12} + n_{21})}{n}. \qquad (7)$$

The value 1 means the total agreement and 0 means that the diagonal frequencies equal the frequencies in two other cells of the contingency table (this case means independence of the variables according to the odds ratio).

For asymmetric binary variables it is suitable to apply some special measure. For experiments presented in this paper, the *Jaccard coefficient* was calculated according to the formula

$$J = \frac{n_{11}}{n_{11} + n_{12} + n_{21}}, \qquad (8)$$

where $n_{11}$ is a number of occurrences of the investigated category in both variables together.

### 1.2 Evaluation of clustering

For evaluation of result clusters, the measure based on the Gini coefficient is applied. The Gini coefficient is determined for variability measurement of nominal variables (sometimes called mutability). It sums squared relative frequencies and the sum is subtracted from 1. This coefficient can be divided by the maximum possible value for the purpose of obtaining the value from the interval $\langle 0; 1 \rangle$. For $m$ variables and $n$ objects and clustering of variables into $k$ clusters, the *normalized within-cluster mutability* is expressed as

$$WCM(k) = \frac{1}{n}\frac{K}{K-1}\sum_{g=1}^{k}\frac{m_g}{m}\sum_{i=1}^{n}\left(1-\sum_{u=1}^{K}\left(\frac{n_{giu}}{m_g}\right)^2\right), \tag{9}$$

where $m_g$ is a number of variables in the $g$th cluster and $n_{giu}$ is a frequency of the $u$th category for the $i$th object in the $g$th cluster. This WCM coefficient is based on the $G'$ measure, which was proposed by Řezanková et al. (2011) for the purpose of evaluation of object clustering.

## 2 Experiments

To investigate the influence of applied similarity measures for the assignment variables to clusters, two sets of variables from a sociological research were selected. The data file was obtained from the archives of the Institute of Sociology (IS) of the Academy of Sciences of the Czech Republic[1]. The research is called *Men and Women with a University Degree* (research number 0136). The author of the research is team Gender in Sociology of IS. The survey was realized in 1998 by Sofres-Factum, Prague.

### 2.1 Description of sets of variables

The first group of variables concerns opinions of the respondents on opportunities of men and women in their job. There are 10 variables with categories: women have better opportunities than men, women have the same opportunities as men, and women have worse opportunities than men. The content of these variables is to succeed, to get a job, to have a higher salary for the same work, to get a post of head, to be a director, to advance to higher positions, to increase earnings, to get remunerations, to have authority, and to keep a job. The cases with missing values were omitted, so answers of 1,886 respondents were analyzed.

The second group of variables investigates whether a graduate made a certain decision for family reasons. There are 9 variables with two categories (yes and no). The questions concern part-time work, shift work, flextime, change of job, change of profession, moving, non-use of interesting job offers, refusal of an offer for a higher position, cheat work. The cases with missing values were also omitted; the dataset with 1,904 cases was analyzed.

Concerning a frequency distribution of categories, the first category occurs rarely (only from 1.1 to 2.5 %) in the first data set. The frequencies of the second category are from 40.8 to 82.3 % and for the third category it is from 15.2 to 57.8 %. In the second data set, the category “yes” is present from 8.5 to 25.9 % (category “no” from 74.1 to 91.5 %).

[1] *Czech Social Science Data Archive* (http://archiv.soc.cas.cz/)

## 2.2 Evaluation of clustering

Evaluation of clustering for different numbers of clusters (from 2 to 5) is presented in Tables 1 and 2, which contain values of the WCM measure. The applied complete linkage method gives better results than other approaches of hierarchical cluster analysis. The single and average methods were also used; the results were worse or the same in most cases.

**Tab. 1: Evaluation of clustering of three-category variables (complete linkage method)**

| | *WCM*(2) | *WCM*(3) | *WCM*(4) | *WCM*(5) |
|---|---|---|---|---|
| Cramer's V | 0.416 | 0.354 | 0.287 | 0.208 |
| Uncertainty coefficient | 0.427 | 0.352 | 0.287 | 0.208 |
| Coefficient kappa | 0.394 | 0.337 | 0.276 | 0.209 |
| Overlap | 0.381 | 0.321 | 0.260 | 0.195 |

Source: data from research *Men and Women with a University Degree*, own calculations

**Tab. 2: Evaluation of clustering of two-category variables (complete linkage method)**

| | *WCM*(2) | *WCM*(3) | *WCM*(4) | *WCM*(5) |
|---|---|---|---|---|
| Cramer's V | 0.366 | 0.320 | 0.255 | 0.186 |
| Uncertainty coefficient | 0.366 | 0.320 | 0.254 | 0.186 |
| Hamann coefficient & overlap | 0.366 | 0.301 | 0.236 | 0.172 |
| Jaccard coefficient | 0.402 | 0.320 | 0.250 | 0.186 |

Source: data from research *Men and Women with a University Degree*, own calculations

From Tables 1 and 2 we can see that the best results (the smallest variability within clusters) were obtained when using the overlap measure. In case of the two-category variables, the obtained values of the WCM coefficient for this measure equal those when the Hamann coefficient (as a measure of agreement) was applied. For the three-category variables, clustering with the kappa coefficient is at the second position. The Jaccard coefficient applied to clustering of two-categories variables gave a little worse results but this measure is important for taking the asymmetric binary variables into account. For the three and five-cluster solutions the assignment of variables to clusters obtained by this coefficient is the same as those obtained using the association measures (Cramer's V and the uncertainty coefficient). For the four-cluster solution the result obtained by the Jaccard coefficient is better than results given by the association measures.

With regard to substantive interpretation, the most dependent variables are also the most similar according to the measures of agreement and the overlap measure. There are variables *to get a post of head* and *to be a director* in the three-category set and variables *change of job* and *change of profession* in the two-category set. Three-category variables *to increase earnings* and *to get remunerations* are also highly similar according to all measures. During clustering process one gets many various solutions. In both sets the most similar variables are in a separated cluster until the five-cluster solution by the overlap measure. In the three-category dataset there is the second pair with other measures. To make the final decision concerning the applied coefficients, many data sets should be analyzed.

## Conclusion

Clustering of nominal variables using different similarity measures was investigated in this paper. If variables have the same number of categories with the same meaning, besides of association measures, the measures of agreement and measures for object similarity evaluation can be applied. The analyses of two data sets show that in such cases the measures of agreement and the simple matching coefficient (the overlap measure) give better results than measures of association taking into account the within-cluster variability. It would be useful to investigate some other measures for object similarity evaluation which have been proposed recently, see e.g. the paper (Šulc, 2014), which compares some of them.

The datasets often contain variables with categories with different meaning. In this case, there is no other possibility than to use the measures of association. If asymmetric binary variables are clustered, special measures for this type of variables should be applied and different approach for the final evaluation should be proposed. Similarly, if ordinal variables should be clustered, special measures taking order of categories into account have to be used.

It is also useful to determine the optimal number of clusters. Using the WCM coefficient proposed in this paper, we can only evaluated the quality of clustering. The values of this measure decrease with the increasing number of clusters. Approaches inspired by coefficients proposed for quantitative variables, see e.g. (Gan et al., 2007; Löster and Pavelka, 2013), would be useful. Some have been proposed for evaluation of object clustering when objects are characterized by nominal variables, see (Řezanková et al., 2011). They could be also used for evaluation of nominal variable clustering.

## Acknowledgment

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

## References

Anderberg, M. R. (1973). *Cluster Analysis for Applications*. New York: Academic Press.

Bohdalová, M. & Greguš, M. (2011). PCA Factor model for forward euro exposure. In Löster, T., Pavelka, T. (Eds.), *International Days of Statistics and Economics*. Slaný: Melandrium, 61-71. ISBN 978-80-86175-77-5.

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. Berlin Heidelberg: Springer, 91-99.

Frolov, A. A., Húsek, D., Polyakov, P. Y. (2014). Two expectation-maximization algorithms for Boolean factor analysis. *Neurocomputing*, *130*(SI), 83-97.

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia: ASA-SIAM.

Li, S., Vel, O., & Coomans, D. (1995). *Comparative performance analysis of nonlinear dimensionality reduction methods*. North Australia: James Cook University.

Löster, T. & Pavelka, T. (2013). Evaluating of the results of clustering in practical economic tasks. In Löster, T., Pavelka, T. (Eds.), *International Days of Statistics and Economics*. Slaný: Melandrium, 804–818. ISBN 978-80-86175-87-4.

Maaten, L. P. J., Postma, L. O., & Herik, H. J. (2008). *Dimensionality reduction: a comparative review*. Elsevier.

Prokop, M. & Řezanková, H. (2011). Data dimensionality reduction methods for ordinal data. In Löster, T., Pavelka, T. (Eds.), *International Days of Statistics and Economics*. Slaný: Melandrium, 523-533. ISBN 978-80-86175-77-5.

Řezanková, H. & Löster, T. (2013). Cluster analysis of households characterized by categorical indicators (in Czech). *E+M. Ekonomie a Management, 16*(3), 139-147.

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*. Berlin: Springer Verlag, 173-182. ISBN 978-3-642-18028-6.

Šulc, Z. (2014). Comparison of the new approaches in nominal data clustering (in Czech). In *Sborník prací vědeckého semináře doktorského studia FIS VŠE*. Praha: Oeconomica, 214-223. ISBN 978-80-245-2010-0.

## Contact

Hana Řezanková
University of Economics, Prague, Dept. of Statistics and Probability
W. Churchill sq. 4, 130 67 Prague 3, Czech Republic
hana.rezankova@vse.cz