NOMCLUST: AN R PACKAGE FOR HIERARCHICAL CLUSTERING OF OBJECTS CHARACTERIZED BY NOMINAL VARIABLES

Zdeněk Šulc – Hana Řezanková

Abstract

Clustering of objects characterized by nominal (categorical) variables is not still sufficiently solved in software systems. Most of them provide only hierarchical cluster analysis with one basic similarity measure for nominal data, the simple matching coefficient, which provides worse results than clustering based on some of recently proposed measures. In this paper, we introduce the *nomclust* R package, which completely covers hierarchical clustering of objects characterized by nominal variables from a proximity matrix computation to final clusters evaluation. It enables to choose among 11 similarity measures, three linkage methods of hierarchical clustering, and uses six evaluation criteria based on the mutability and the entropy. Some of the criteria were constructed for comparison of different cluster approaches and some for determining the optimal number of clusters. The package contains both the main function, which covers the whole clustering process, and a set of subsidiary functions, which allow computing only a part of clustering process, e.g. a proximity matrix, or evaluation of given cluster solution. We illustrate the methodology of this package on a small dataset.

Key words: cluster analysis, nominal variables, similarity measures, evaluation criteria

JEL Code: C38, C88

Introduction

In this paper, we present the *nomclust* package for the R software, which was developed for hierarchical clustering of objects characterized by nominal variables. This package is available on the Comprehensive R Archive Network (CRAN) web site[1]. We developed it due to the lack of software implementation of similarity measures determined for nominal variables. Nominal variables are currently clustered by means of methods based on the simple matching (SM) coefficient, which are usually used in statistical software. This treatment is not

[1]http://CRAN.R-project.org/

sufficient though, because it neglects important characteristics of a dataset, such as frequencies of categories or number of categories, which could be used for better similarity determination. The other possibility is to use measures determined for mixed type data. One of the famous ones is the *Gower similarity coefficient*, see (Gower, 1971), which can be found e.g. in function *daisy* in the R package *cluster*, or the *log-likelihood distance* in *two-step cluster analysis*, see (Chiu et al., 2001), in the IBM SPSS software. However, the Gower coefficient treats nominal variables in the same way as measures based on the SM coefficient, and as far as we know, the log-likelihood distance is not available in any non-commercial software. Another possibility is to transform the original nominal variables into a set of dummy variables, to which measures for quantitative or binary data can be applied. However, this procedure always provides some loss of information caused by dummy transformation. Overall, current software solutions do not provide satisfactory results.

In recent years, there have been introduced many similarity measures which were proposed for nominal variables, e.g. the Lin measure, see (Lin, 1998), or the Eskin measure, see (Eskin et al., 2002). Many of them provide very good results as well, see (Šulc and Řezanková, 2014) or (Šulc, 2015). The problem is that those measures have not become widely known, mostly due to the lack of software implementation. By introducing our R package, we aspire to deal with this problem. The package contains 11 similarity measures determined for nominal data, and completely covers the clustering process – from proximity matrix computation, over hierarchical clustering analysis, to evaluation criteria of resulting clusters determination. In this paper, we introduce the functionality of this package in detail. The paper is organized as follows. The Section 1 offers a short introduction of used similarity measures, linkage methods, and evaluation criteria. The Section 2 presents several typical use illustrations on a simple example. Final remarks are placed in Conclusion.

1 Theoretical background of the package

The package contains 11 similarity measures for nominal data, 10 of them were summarized in (Boriah et al., 2008), and one of them was introduced by Morlini and Zani, (2012). All the measures were more thoroughly examined in (Šulc and Řezanková, 2014) or (Šulc, 2015).

Tab. 1 presents equations which determine similarity between the i-th and j-th objects by the c-th variable. This is the first (and the most important) level of a proximity matrix computation. On the second level, similarity for objects across all variables is computed. The last level involve computation of the dissimilarity. All formulas in Tab. 1 are based on the

data matrix $\mathbf{X} = [x_{ic}]$, where $i = 1, 2, ..., n$ (n is the total number of objects); $c = 1, 2, ..., m$ (m is the total number of variables). The number of categories of the c-th variable is denoted as n_c, absolute frequency as f, and relative frequency as p.

Tab. 1: Similarity measures in the *nomclust* package(without dummy transformation)

Measure	Similarity between the i-th and j-th objects by the c-th variable
Eskin	$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ if } x_{ic} = x_{jc} \\ \dfrac{n_c^2}{n_c^2 + 2} \text{ otherwise} \end{cases}$
Goodall 1	$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 - \sum_{q \in Q} p_c^2(q) \text{ if } x_{ic} = x_{jc} \\ 0 \text{ otherwise} \end{cases}, Q \subseteq X_c : \forall q \in Q, p_c(q) \le p_c(x_{ic})$
Goodall 2	$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 - \sum_{q \in Q} p_c^2(q) \text{ if } x_{ic} = x_{jc} \\ 0 \text{ otherwise} \end{cases}, Q \subseteq X_c : \forall q \in Q, p_c(q) \ge p_c(x_{ic})$
Goodall 3	$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 - p_c^2(x_{ic}) \text{ if } x_{ic} = x_{jc} \\ 0 \text{ otherwise} \end{cases}$
Goodall 4	$S_c(x_{ic}, x_{jc}) = \begin{cases} p_c^2(x_{ic}) \text{ if } x_{ic} = x_{jc} \\ 0 \text{ otherwise} \end{cases}$
IOF	$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ if } x_{ic} = x_{jc} \\ \dfrac{1}{1 + \ln f(x_{ic}) \cdot \ln f(x_{jc})} \text{ otherwise} \end{cases}$
Lin	$S_c(x_{ic}, x_{jc}) = \begin{cases} 2 \cdot \ln p(x_{ic}) \text{ if } x_{ic} = x_{jc} \\ 2 \cdot \ln(p(x_{ic}) + p(x_{jc})) \text{ otherwise} \end{cases}$
Lin 1	$S_c(x_{ic}, x_{jc}) = \begin{cases} \sum_{q \in Q} \ln p_c(q) \text{ if } x_{ic} = x_{jc} \\ 2 \ln \sum_{q \in Q} p_c(q) \text{ otherwise} \end{cases}, Q \subseteq X_c : \forall q \in Q, p_c(x_{ic}) \le p_c(q) \le p_c(x_{jc})$
OF	$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ if } x_{ic} = x_{jc} \\ \dfrac{1}{1 + \ln \dfrac{n}{f(x_{ic})} \cdot \ln \dfrac{n}{f(x_{jc})}} \text{ otherwise} \end{cases}$
SM	$S_c(x_{ic}, x_{jc}) = \begin{cases} 1 \text{ if } x_{ic} = x_{jc} \\ 0 \text{ otherwise} \end{cases}$

The *Morlini and Zani's* similarity measure uses dummy transformation; thus, it is computed in a different way than the rest of measures. Similarity between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$ is expressed:

$$S(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{c=1}^{m}\sum_{u=1}^{K_c} \tau(i,j)_{cu} \ln\left(\frac{1}{f_{cu}^2}\right)}{\sum_{c=1}^{m}\sum_{u=1}^{K_c} \tau(i,j)_{cu} \ln\left(\frac{1}{f_{cu}^2}\right) + \sum_{c=1}^{m} \phi(i,j)_c \sum_{u=1}^{K_c} f_{cu} \ln\left(\frac{1}{f_{cu}^2}\right)}, \quad (1)$$

Where $u = 1, 2, \ldots K_c$ is an index of the u-th dummy variable of the c-th original variable, f_{cu}^2 is the second power of the frequency of the u-th category of the original c-th variable. When using Eq. (1), the following conditions have to be held:

$$\tau(i,j)_{cu} = 1 \text{ if } x_{icu} = 1 \text{ and } x_{jcu} = 1,$$
$$\tau(i,j)_{cu} = 0 \text{ otherwise,}$$

$$\phi(i,j)_c = 1 \text{ if } x_{icu} = 1 \cap x_{jct} = 1 \text{ and } x_{ict} = 1 \cap x_{jcu} = 0, \text{ for } u \neq t,$$
$$\phi(i,j)_c = 0 \text{ otherwise.}$$

After computation of proximity matrix, hierarchical cluster analysis is performed. The package enables to use three different linkage methods – complete, single and average one. They differ from the way, how they determine dissimilarity between two clusters. The complete linkage method consider this dissimilarity as dissimilarity between two furthest objects from two different clusters. The single linkage method uses dissimilarity between two closest objects, and the average linkage method takes average pairwise dissimilarity between objects in two different clusters.

The resulting clusters can be evaluated by six criteria, which were introduced in (Řezanková et al., 2011). They are based on the within-cluster variability of clusters, which can be measured either by the mutability or by the entropy. The criteria are presented in Tab. 2, where k is the number of clusters, n_g is the number of objects in the g-th cluster ($g = 1, 2, \ldots, k$), n_{giu} is the number of objects in the g-th cluster by the i-th object with the u-th category ($u = 1, 2, \ldots, K_c$; K_c is the number of categories), $c = 1, 2, \ldots, m$ (m is the number of variables).

WCM and WCE measure the within-cluster variability in a dataset. They take values from zero to one, where values close to one indicate high variability. With increasing number of clusters, the within-cluster variability decreases, and thus, values of these indices decrease.

Values of coefficients PSTau and PSU take values from zero to one as well. This time, values close to one indicate the low within-cluster variability and vice versa. Pseudo F coefficients, originally determined for quantitative data, see e.g. (Löster, 2014), are based on the mutability or the entropy for nominal variables. They were developed to determine the optimal cluster solution, i.e. the solution with the relative highest decrease of within-cluster variability. Maximal value should indicate the best solution.

Tab. 2: Evaluation criteria in the *nomclust* package

Within-cluster mutability coefficient (WCM)	$WCM(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \frac{K_c}{K_c - 1} \left(1 - \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \right)^2 \right)$
Within-cluster entropy coefficient (WCE)	$WCE(k) = \sum_{g=1}^{k} \frac{n_g}{n \cdot m} \sum_{c=1}^{m} \frac{1}{\ln K_c} \left(- \sum_{u=1}^{K_c} \left(\frac{n_{gcu}}{n_g} \ln \frac{n_{gcu}}{n_g} \right) \right)$
Pseudo tau coefficient (PSTau)	$PSTau(k) = \frac{WCM(1) - WCM(k)}{WCM(1)}$
Pseudo uncertainty coefficient (PSU)	$PSU(k) = \frac{WCE(1) - WCE(k)}{WCE(1)}$
Pseudo F coefficient based on the mutability (PSFM)	$PSFM(k) = \frac{(n-k)(WCM(1) - WCM(k))}{(k-1)WCM(k)}$
Pseudo F coefficient based on the entropy (PSFE)	$PSFE(k) = \frac{(n-k)(WCE(1) - WCE(k))}{(k-1)WCE(k)}$

Coefficients based on the mutability and the entropy usually do not differ very much. They are included in the package in order to provide two independent ways of variability computation. Big differences between mutability- and entropy-based coefficients should attract researcher's attention. The use of two types of coefficients is especially useful when determining the optimal number of clusters using the pseudo F coefficients. If both the coefficients prefer the same cluster solution, there is strong evidence for choosing this particular solution by a researcher.

2 Illustrations of use

There are three possible ways how to use the *nomclust* package. One can use it to a proximity matrix computation using one of 11 available similarity measures, or to perform the complete

clustering process, where the input is data matrix **X**, and the output consists of cluster membership variables, and a set of evaluation criteria. The third way of the use is to compute a set of evaluation criteria of a dataset with cluster membership variables, which were obtained by any other clustering process. In this section, all the ways of use will be described in detail. Examples will be demonstrated on the dataset *data20*, which contains five nominal variables, each with 20 observations. It is available in the package after typing:

```
R> data(data20)
```

2.1 Proximity matrix computation

In case one wants to compute a proximity matrix using only one of available similarity measures, e.g. for research purposes, the best way is to use directly the function of a particular measure. All the functional calls are presented in Tab. 3. The resulting proximity matrices have standard proportions; their size is *n*x*n*, and they have zero diagonal elements. Thus, they can be used as the input in other packages as well, for instance *hclust* in the *stats* package.

Tab. 3: Functional calls for proximity matrices computation

Measure	Functional call	Measure	Functional call
Eskin	`R>eskin(data)`	Morlini	`R>morlini(data)`
Goodall 1	`R> good1(data)`	Lin	`R> lin(data)`
Goodall 2	`R> good2(data)`	Lin 1	`R> lin1(data)`
Goodall 3	`R> good3(data)`	OF	`R> of(data)`
Goodall 4	`R> good4(data)`	SM	`R>sm(data)`
IOF	`R>iof(data)`		

2.2 Complete clustering process

The most complex function in the package is the *nomclust* function. It covers the whole clustering process, and it has following syntax:

```
R>nomclust(data, measure = "iof", clu_low = 2, clu_high = 6,
eval = TRUE, prox = FALSE, method = "complete")
```

The only mandatory parameter is *data*, which represents the input data for the analysis. The rest of parameters is preset by default in a way which should suit to majority of users, but it can be changed if needed. Parameter *measure* chooses the similarity measure for the analysis.

As a default choice, the IOF measure is chosen, because it often provides the best results, see (Šulc and Řezanková, 2014). Generally, any of measures from Tab. 3 can be used in its place. Parameters *clu_low* and *clu_high* determine the lower and upper bounds for the number of created cluster solutions. There are two logical parameters *eval* and *prox*, which enable to compute evaluation criteria or to display a proximity matrix respectively. The last parameter is *method*, which enables to use one of three linkage methods. By default, the complete linkage method is chosen, because it usually provides the best results; see (Šulc, 2014). In the package, the *agnes* algorithm from the package *cluster* was used for performing the hierarchical cluster analysis. In order to use the *nomclust* function on the *data20* dataset with the IOF measure, and displayed proximity matrix, the following syntax is needed:

```
R> data <- nomclust(data20, prox = TRUE)
```

The output of the *nomclust* function has a form of a list with up to three components. The *mem* component contains cluster memberships for all cases. Because the output would be large, only first six rows of the output are displayed:

```
R>clu_mem<- head(data$mem)
```

```
   clu_2 clu_3 clu_4 clu_5 clu_6
1      1     1     1     1     1
2      1     1     1     1     1
3      1     2     2     2     2
4      1     1     1     3     3
5      1     1     1     1     1
6      2     3     3     4     4
```

The evaluation criteria can be found in the *eval* component after typing:

```
R>clu_eval<- data$eval
```

```
   cluster  WCM     WCE     PSTau   PSU     PSFM    PSFE
1        1  0.9666  0.9600     NA      NA      NA      NA
2        2  0.7330  0.7010  0.1987  0.2084  4.4646  4.7382
3        3  0.6127  0.5789  0.3365  0.3607  4.3106  4.7949
4        4  0.4546  0.4066  0.5005  0.5491  5.3446  6.4961
```

```
5         5   0.4136   0.3658   0.5373   0.5807   4.3551   5.1943
6         6   0.3600   0.3085   0.6004   0.6514   4.2074   5.2327
```

The output contains values for all the evaluation criteria presented in Tab. 2 for the defined range of clusters. In the first row, there is always displayed variability in the whole dataset by means of the WCM and WCE coefficients. The other evaluation criteria have always symbol NA in this row. The other rows represent the within-cluster variability decrease pretty well. When studying the PSFM and PSFE coefficients, both of them prefer the four-cluster solution.

The *prox* component displays proximity matrix, if the logical parameter is set to TRUE. Since the proximity matrix would be large, the output is omitted.

```
R>clu_prox<- data$prox
```

2.3 Clustering evaluation

Sometimes it may arise a need to evaluate set of cluster solutions obtained by a similarity measure or method which is not available in the *nomclust* package, for instance, results obtained by two-step cluster analysis in IBM SPSS. For such cases, one can use the *evalclust* function, which has the following syntax:

```
R>evalclust(data, num_var, clu_low = 2, clu_high = 6)
```

The function has two mandatory parameters; *data*, which is an original dataset with cluster membership variables in an increasing order, and *num_var*, which defines the number of original variables in the dataset. Again, *clu_low* and *clu_high* parameters define the range of used cluster solutions. Thus, evaluation criteria computation for the *data20* dataset with five additional cluster membership variables (2 to 6 clusters) can be written in the following way:

```
R>evalclust(data20, 5, clu_low = 2, clu_high = 6)
```

```
   cluster  WCM      WCE      PSTau    PSE      PSFM     PSFE
1        1   0.9666   0.9600       NA       NA       NA       NA
2        2   0.7601   0.7214   0.1943   0.2366   4.3400   5.5802
3        3   0.6118   0.5670   0.3490   0.4034   4.5570   5.7473
4        4   0.4615   0.4020   0.4851   0.5488   5.0252   6.4857
5        5   0.3938   0.3284   0.5662   0.6355   4.8955   6.5379
```

```
6         6   0.3283   0.2649   0.6293   0.6941   4.7540   6.3542
```

The output has the same form as component *eval* in the *nomclust* function. Thus, a direct comparison of outputs is ensured. For instance, in this particular case, two-step cluster analysis provides comparable results to those ones gotten by the IOF measure used in the *nomclust* function. Moreover, PSFM and PSFE determine the same optimal solution in case of IOF, whereas their values in case of two-step cluster analysis evaluation differ. For more information about two-step cluster analysis clustering performance, see e.g. (Šulc and Řezanková, 2014).

Conclusion

In this paper, we introduced the main features of the *nomclust* package. We think it offers a comprehensive look at the nominal clustering issue which cannot be found in any other R packages or commercial software. We are aware that the R package development is a living process; and thus, we are going to improve it steadily.

In next releases, we would like to focus on performance optimization, because proximity matrices computation takes a lot of computational time by large datasets. Next, we will incorporate graphical outputs of the analysis, which should make a cluster evaluation process even easier.

Acknowledgments

This paper was processed with contribution of long term institutional support of research activities by Faculty of Informatics and Statistics, University of Economics, Prague.

References

Boriah, S., Chandola, V., Kumar, V. (2008). Similarity measures for categorical data: a comparative evaluation. In *Proceedings of the 8th SIAM International Conference on Data Mining*. SIAM, 243-254.

Chiu, T., Fang, D., Chen, J., Wang, Y., Jeris, C. (2001). A robust and scalable clustering algorithm for mixed type attributes in large database environment. In *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, 263.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. In *Biometrics*, *28*(4), 857-871.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., Stolfo, S. (2002). A geometric framework for unsupervised anomaly detection. In *Applications of Data Mining in Computer Security*, 78-100.

Morlini, I., Zani, S. (2012). A new class of weighted similarity indices using polytomous variables. In *Journal of Classification*, *29*(2), 199-226.

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., 296-304.

Löster, T. (2014). The evaluation of CHF coefficient in determining the number of clusters using Euclidean distance measure. In *The 8th International Days of Statistics and Economics*. Slaný: Melandrium, 2014, 858-869. http://msed.vse.cz/msed_2014/article/463-Loster-Tomas-paper.pdf.

Řezanková, H., Löster, T.,Húsek, D. (2011). Evaluation of categorical data clustering. In Mugellini, E, Szczepaniak, P. S., Pettenati, M. C. et al. (Eds.), In *Advances in Intelligent Web Mastering 3*. Berlin: Springer Verlag, 173-182.

Šulc, Z. (2014). Similarity Measures for Nominal Variable Clustering. In *The 8th International Days of Statistics and Economics*. Slaný: Melandrium, 1536-1545. http://msed.vse.cz/msed_2014/article/275-Sulc-Zdenek-paper.pdf.

Šulc, Z. (2015). Application of Goodall's and Lin's similarity measures in hierarchical clustering. In *Sborník prací vědeckého semináře doktorského studia FIS VŠE*. Praha: Oeconomica, 112-118.
http://fis.vse.cz/wp-content/uploads/2015/01/DD_FIS_2015_CELY_SBORNIK.pdf.

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*. Wroclaw: Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu, 249-258.

Zdeněk Šulc

University of Economics, Prague, Dept. of Statistics and Probability

W. Churchill sq. 4, 130 67 Prague 3, Czech Republic

zdenek.sulc@vse.cz

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