Developments in Statistics
Andrej Mrvar and Anuška Ferligoj (Editors)
Metodološki zvezki, 17, Ljubljana: FDV, 2002

Symbolic Data Analysis Approach to Clustering Large Ego-Centered Networks

Simona Korenjak-Černe[1]

Abstract

In the paper an adapted version of the leaders clustering method as an efficient method for clustering ego-centered networks is presented. The original data are transformed into symbolic objects. Clustering problem is defined as an optimization problem. Based on this definition an adapted leaders method was developed. The example on the dataset on social support in Ljubljana collected by the Faculty of social sciences at the University of Ljubljana is presented. The clustering of the data was done using the adapted leaders program from the program package CLAMIX.

1 Introduction

Clustering is the grouping of similar objects (Hartigan, 1975). Clusters in the clustering are usually described with a central object (described by mean values of its variables). For better and more detailed descriptions of the objects and their clusters Diday introduced new descriptions called ***symbolic objects*** (Diday, 1979; Diday, 1997).

In the social science research object of interest – units are mostly respondents. Beside respondents (***ego***s) researches are frequently interested also for their personal networks (***alter***s). In this case we are talking about ***ego-centered*** or ***personal*** networks. Such networks are usually presented in two datasets: one for egos and one for ego's alters.

Every ego is described by selected variables. Alters can be described with the same variables, but usually several (other) variables are also measured on them, which describe relationship among ego and alters and properties of alters (Müller, Wellman, and Marin, 1999).

Three important problems related encountered during the analysis of a set of personal networks are considered in the paper:

[1] Faculty of Economics, University of Ljubljana, Slovenia; Simona.Cerne@uni-lj.si.

- the problem of ***units description***: the units for analysis are descriptions of egos and their personal networks. In such descriptions we want to save as much information about egos and their alters as possible;
- the problem of ***data size***: the (random) sample of egos can be large – some hundreds of egos and the corresponding set of personal networks consists of some thousands of alters;
- the problem of ***mixed units***: variables used for describing egos and alters are usually measured in different scales (nominal, ordinal, interval or ratio).

To describe egos and their alters we shall use a special type of symbolic objects (Diday, 1997) using variables distributions. They provide a more detailed description of personal networks than the 'standard' approach where descriptions with the value of an appropriate statistics – e.g., mean value are used.

The size of the dataset is reduced with clustering method, where units are clustered into selected number of groups (clusters) in which the units are as similar as possible. For large datasets, hierarchical clustering methods are not the best choice for this task. These methods are mainly based on the dissimilarity matrix and are therefore of higher than quadratic complexity. Because of this they are not appropriate for large datasets.

For large datasets, usually the local optimization clustering methods are used (e.g., k-means method). The main drawback of these methods is that they are implemented only for numerical data. Korenjak-Černe and Batagelj (1998) proposed an adapted version of the leaders clustering method that can efficiently cluster ***large*** datasets with units described with ***symbolic objects***.

Although we shall describe in details the clustering of units, several other methods of symbolic data analysis can also be applied to symbolic descriptions of personal networks (Bock and Diday, 2000).

2 Symbolic object description of a personal network

Let denote the finite set of egos with $\mathbf{E}$ and with $\mathbf{A}$ the set of alters. To each ego $\mathbf{X} \in \mathbf{E}$ corresponds a group of alters $\mathbf{A}(\mathbf{X})$. It holds $\bigcup_{\mathbf{X}\in\mathbf{E}} \mathbf{A}(\mathbf{X}) = \mathbf{A}$. Suppose that each ego $\mathbf{X} \in \mathbf{E}$ is described with m_e variables U_1, U_2, U_3, $\ldots$, U_{m_e}

$$\boldsymbol{X}^o = [u_1, u_2, u_3, \ldots, u_{m_e}]$$

where $u_i = U_i(\mathbf{X})$ is the value of the i-th variable U_i on ego $\mathbf{X}$; and that each alter $\mathbf{Y}$ is described with m_a variables V_1, V_2, V_3, $\ldots$, V_{m_a}

$$\boldsymbol{Y}^o = [v_1, v_2, v_3, \ldots, v_{m_a}]$$

where $v_i = V_i(\mathbf{Y})$.

We represent an ego's alters group with a symbolic object $So(\mathbf{A}(\mathbf{X}))$ that consists of distributions f_j of its alter's values for each variable V_j.

$$So(\mathbf{A}(\mathbf{X})) = [\mathbf{f_1}, \mathbf{f_2}, \mathbf{f_3}, \ldots, \mathbf{f_{m_a}}]$$

$$f_j \equiv f(\mathbf{X}; \mathbf{V_j}) = [\mathbf{f(1, X; V_j), f(2, X; V_j), f(3, X; V_j)}, \ldots, \mathbf{f(k_j, X; V_j)}]$$

They are constructed as follows. The domain of the variable V is partitioned into k_V subsets. Then f_i of the respondent (ego) $\mathbf{X}$ is determined by the relative frequencies

$$f(i, \mathbf{X}; \mathbf{V}) \quad = \quad \frac{q(i, \mathbf{X}; \mathbf{V})}{\mathrm{card}(\mathbf{A}(\mathbf{X}))} \qquad \text{for} \quad i = 1, \ldots k_V$$

where $q(i, \mathbf{X}; \mathbf{V}) = \mathrm{card}(\mathbf{Q}(\mathbf{i}, \mathbf{X}; \mathbf{V}))$ and

$$Q(i, \mathbf{X}; \mathbf{V}) \quad = \quad \{\mathbf{Y} : \mathbf{Y} \text{ is an alter of the ego } \mathbf{X} \text{ with the value in the } i\text{-th subset}\}.$$

With $\mathrm{card}(C)$ the number of elements in the set C is denoted.
Finally we combine the description of ego $\mathbf{X}$ and symbolic object describing its alters into an extended ego's description – symbolic object

$$So(\mathrm{X}) = [\boldsymbol{X}, \boldsymbol{f}\,]$$

These symbolic objects are the basis for symbolic data analysis of personal networks. They provide an ***uniform*** description of originally ***mixed*** units and their clusters.

2.1 An example of a symbolic object description of an ego

For three selected variables of egos measured in different scales: U_1 – *Satisfied with informational support* (ordinal), U_2 – *Average monthly income (in SIT)* (numerical - ratio), and U_3 – *Sex* (nominal), we use the following partitions of their domains:

$U_1 =$ ***Satisfied with informational support***
1 = very dissatisfied
2 = quite dissatisfied
3 = little dissatisfied
4 = little satisfied
5 = quite satisfied
6 = very satisfied

$U_3 =$ ***Sex***
1 = male
2 = female

$U_2 =$ ***Average monthly income***
1 = less than 50.000 SIT
2 = 50.000 SIT
3 = between 50.000 and 100.000 SIT
4 = 100.000 SIT
5 = between 100.000 and 150.000 SIT
6 = 150.000 SIT
7 = between 150.000 and 200.000 SIT
8 = 200.000 SIT
9 = more than 200.000 SIT
10 = no income

and for two selected variables of alters: $V_1 =$ *Sex of the alter* (nominal) and
$V_2 =$ *Frequency of contact* (ordinal)

$V_1 =$ ***Sex*** of the alter
1 = male
2 = female

$V_2 =$ ***Frequency of contact***
1 = every day
2 = couple of times in a week
3 = couple of times in a month
4 = approximately once a month
5 = couple of times in a year
6 = once in a year or less

The description of selected ego $\mathbf{X_{1001}}$ is

$$\begin{gathered}\boldsymbol{X}_{1001} = [5,4,1] = \\ = [[0,0,0,0,1,0],[0,0,0,1,0,0,0,0,0,0],[1,0]]\end{gathered}$$

From this description it can be seen that selected ego $\mathbf{X_{1001}}$ is 'quite satisfied with alters informational support', is 'male' with average monthly income '100.000 SIT'.

The descriptions of three of his alters are

$$\begin{aligned}\boldsymbol{Y}_{1001:01} &= [2,1] \\ &= [[0,1],[1,0,0,0,0,0]] \\ \boldsymbol{Y}_{1001:02} &= [2,3] \\ &= [[0,1],[0,0,1,0,0,0]] \\ \boldsymbol{Y}_{1001:03} &= [1,1] \\ &= [[1,0],[1,0,0,0,0,0]]\end{aligned}$$

Therefore the ego's alters group $\mathbf{A}(\mathbf{X_{1001}}) = \{\boldsymbol{Y}_{\mathbf{1001:01}}, \boldsymbol{Y}_{\mathbf{1001:02}}, \boldsymbol{Y}_{\mathbf{1001:03}}\}$ is described with

$$So(\mathbf{A}(\mathbf{X_{1001}})) = [[\frac{1}{3},\frac{2}{3}],[\frac{2}{3},0,\frac{1}{3},0,0,0]]$$

because one of the alters is man in two are women, and two of them are 'every day' in contact with the ego and one contact with the ego only 'couple of times in a month'.

Finally the combined symbolic object description for the ego $\mathbf{X_{1001}}$ is

$$So(\mathbf{X_{1001}}) = [5,4,1,[\frac{1}{3},\frac{2}{3}],[\frac{2}{3},0,\frac{1}{3},0,0,0]]$$

Such a description has the following important properties:

- it requires a ***fixed space*** per variable;
- it is ***compatible with merging*** of clusters – knowing the description of two disjoint clusters C_1 and C_2, $C_1 \cap C_2 = \emptyset$, we can, without additional information, produce the description of their union
$$f(i, C_1 \cup C_2; V) = \frac{\mathrm{card}(C_1)\, f(i, C_1; V) + \mathrm{card}(C_2)\, f(i, C_2; V)}{\mathrm{card}(C_1) + \mathrm{card}(C_2)};$$
- it produces an ***uniform description*** for all the types of variables;
- with the distributions the ***complete information*** about alters ***is stored***.

3 Clustering as an optimization problem

Suppose that units descriptions consist of m selected variables from $So(\mathbf{X})$. To develop a clustering procedure for their analysis first the dissimilarity between two units $\mathbf{X_1}$ and $\mathbf{X_2}$ is defined as the weighted sum of the dissimilarity between them on each variable V

$$d(\mathbf{X_1}, \mathbf{X_2}) = \sum_{\mathbf{j=1}}^{\mathbf{m}} \alpha_{\mathbf{j}}\ \mathbf{d}(\mathbf{X_1}, \mathbf{X_2}; \mathbf{V_j}), \qquad \sum_{\mathbf{j=1}}^{\mathbf{m}} \alpha_{\mathbf{j}} = \mathbf{1}$$

where

$$d_{abs}(\mathbf{X_1}, \mathbf{X_2}; \mathbf{V}) = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\mathbf{i=1}}^{\mathbf{k_V}} |\mathbf{f}(\mathbf{i}, \mathbf{X_1}; \mathbf{V}) - \mathbf{f}(\mathbf{i}, \mathbf{X_2}; \mathbf{V})| \tag{3.1}$$

or

$$d_{sqr}(\mathbf{X_1}, \mathbf{X_2}; \mathbf{V}) = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\mathbf{i=1}}^{\mathbf{k_V}} (\mathbf{f}(\mathbf{i}, \mathbf{X_1}; \mathbf{V}) - \mathbf{f}(\mathbf{i}, \mathbf{X_2}; \mathbf{V}))^{\mathbf{2}}. \tag{3.2}$$

Here, $\alpha_j \geq 0$ $(j = 1, \ldots, m)$ denote weights, which could be equal for all variables or different if we have same information about their importance. Because the clusters are represented in the same way the same definitions for defining the dissimilarity between two clusters are used.

Based on this definition the clustering problem is defined as the following optimization problem: Among clusterings $\mathbf{C}$ from the set of feasible clusterings Φ find a clustering $\mathbf{C}^*$ for which

$$P(\mathbf{C}^*) = \min_{\mathbf{C} \in \mathbf{\Phi}} \mathbf{P}(\mathbf{C}).$$

For the clustering criterion function $P(\mathbf{C})$ its most common form – the sum of cluster errors is used:

$$P(\mathbf{C}) = \sum_{\mathbf{C} \in \mathbf{C}} \mathbf{p}(\mathbf{C}),$$

where

$$p(C) = \sum_{X \in C} d(X, L_C).$$

L_C is the leader (the representative element) of the cluster C.

4 The adapted leaders method

For clustering very large datasets where variables of units are measured in different scales we adapted the leaders method (Korenjak-Černe and Batagelj, 1998), a variant of the dynamic clustering method (Diday, 1979). This version is based on descriptions of clusters with distributions and on the definition of clustering as an optimization problem as it was described in previous section. The proposed method can be shortly described with the following procedure:

```
determine an initial clustering
repeat
   determine leaders of the clusters in the current clustering;
   assign each unit to the nearest new leader – producing a new clustering
until the leaders do not change any more.
```

A special version of this method is also the well-known k-means method, which is appropriate only for numerical variables (Hartigan, 1975).

The leaders are also symbolic objects described with distributions. They are determined so that they minimize the cluster error. It can be proved that for the first selected criterion function P_{abs} with the dissimilarity d_{abs} (3.1) the optimal leaders are determined with the ***maximal frequencies***; and for the second criterion function P_{sqr} with d_{sqr} (3.2) with the ***average distributions***. This provides an easy interpretation of the clustering results.

5 An example: The dataset on social support in Ljubljana

The proposed approach was applied on the dataset on social support in Ljubljana collected by the Faculty of social sciences at the University of Ljubljana.

The data were collected between March and June 2000 by computer assisted telephone interview (CATI) and computer assisted personal interview (CAPI) for a representative sample of 1033 inhabitants of the city of Ljubljana, Slovenia, described with 38 variables. These respondents produced 5849 alters described with 10 variables. Variables are measured in different scales. The sample was made on the basis of the telephone directory of Ljubljana. The details can be found in the paper by Kogovšek, Ferligoj, Coenders, and Saris (2002).

The purpose of this article was not to analyze the data, but to present some possibilities of the symbolic data analysis approach to ego-centered networks. Because of this only basic explanations and the possibilities of the interpretations of the results are given. More detailed results and the CLAMIX program (the adapted version of the leaders method) are available at URL:
`http://www.educa.fmf.uni-lj.si/datana/`

The dataset of egos and the dataset of alters are combined into one dataset with 1033 units – symbolic objects. Each unit represents an ego with his/her personal network. Symbolic object is described with 48 variables (38 egos and 10 alters variables).

For the initial clustering the partition on 20 clusters was randomly selected. The maximal allowed dissimilarity between the unit and the clusters leader was also limited. The clustering procedure was tested with different selections of variables' weights and in all cases leaders stop changing after less than 20 iterations.

The output files of the leaders program contain a lot of information about each of the clusters, such as: the number of units in the cluster; the error of the cluster; the nearest unit to the leader and it's dissimilarity from the leader; the farthest unit and it's dissimilarity from the leader. Optionally the user can get also the frequencies for each variable; values of the characteristic classes (classes with maximal frequencies) for each variable; and the set of all units in the cluster.

5.1 An example of the description of the output cluster C

Selected cluster C contains 45 units. For the variable DQ3A *(social companionship)* the following information is available:
max frequency = 35 (77.78 %), # missing: 0, # no sense: 0

```
q(C;V):    0   0   0   0   10   35
maxL(V): 0   0   0   0    0    1
```

Explanation: In the cluster C are 45 egos. For the variable DQ3A *(social companionship)* none of them has 'missing' or 'no sense' values. Ten egos in the this cluster are 'quite satisfied with alters social companionship', and most (precisely 35, which is 77.78 %) of the egos in this cluster are 'very satisfied with alters social companionship'.

The percentage of the units with values in the last subset among six subsets of the domain of the variable DQ3A *(social companionship)* is rather high, so it can be said that the values in this subset (in this case the value 'very satisfied') are characteristic for the selected cluster C.

From the distribution above can also be seen that if the subset of values is extended to a nearest subset into the subset {quite satisfied, very satisfied}, all units are included in this extended subset. From that can be concluded that all egos inside the selected cluster C are at least 'quite satisfied' with the social companionship from their alters.

Similar interpretations can be made for all variables because for all of them the whole distributions are available.

Based on the above interpretations of characteristics, the modal subsets of the variables' domains show the following characteristic values of the selected cluster C (a variable is included only if maximal frequency is more than 50%):

91.11 % D2*(how many children under* 19 *in the household)* = no
57.78 % D4*(live in the town (in years))* = from 22 to 25
86.67 % D7*(age)* = from 22 to 26
86.67 % D8*(profession)* = university student
91.11 % D9*(education)* = 4-year high school
95.56 % D10*(marital status)* = single
75.56 % SPOL*(sex)* = female

88.89 % DQ1A*(material support)* = very satisfied
80.00 % DQ2A*(informational support)* = very satisfied
77.78 % DQ3A*(social companionship)* = very satisfied
84.44 % DQ4A*(emotional support)* = very satisfied

55.37 % Q12*(alters sex)* = female
52.44 % Q14*(how far he/she lives)* = 15 min or less

As can be seen from this description in the cluster are mainly 'single' persons (95.56 % of units inside this cluster are single), 'with 4-year high school' (91.11 %), 'female' (75.56 %), 'university students' (86.67 %), 'very satisfied with alters support' (variables DQ1A,DQ2A,DQ3A,DQ4A), have friends living close to them (52.44 % of ego's alters live 15 minutes or less from the ego).

In the interpretations also the number of subsets of the variables domain is important. For example, the variable *Sex* has only two possible values and therefore its domain is divided into only two subsets. If the percentage in one of them is only a little more than 50%, we can not say that such value is characteristic. More appropriate interpretation is that the cluster is in such case well balanced regarding to that variable. Based on such interpretation the selected cluster is well balanced regarding to the sex of the egos' alters. But on the opposite, the domain of the variable Q14 *(how far he/she lives)* is partitioned into five subsets so the relative frequency in one of them can be considered as rather high if it is more than 50%.

6 Conclusion

The main advantages of symbolic data analysis approach to ego-centered networks are:

- the analysis is based on descriptions of ego-centered networks with ***distributions***; thus preserving more information than methods based on central values, and providing an uniform description of ***mixed*** units;
- it is appropriate for ***large*** datasets;
- variables can be ***weighted*** by their importance.

Acknowledgment

The author would like to thank her advisor professor Vladimir Batagelj for all his suggestions and help working on clustering problem. Thanks also go to professor Anuška Ferligoj and assistant professor Tina Kogovšek for providing the dataset on social support used for the example and for stimulating discussions about it.

References

[1] Bock, H.H. and Diday, E. (Eds.) (2000): *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data.* Heidelberg: Springer.

[2] Diday, E. (1997): *Extracting Information from Extensive Data sets by Symbolic Data Analysis.* Indo-French Workshop on Symbolic Data Analysis and its Applications, Paris, September 1997, Paris IX, Dauphine, 3-12.

[3] Diday, E. (1979): *Optimisation en Classification Automatique*, Tome 1.,2.., Rocquencourt: INRIA (in French).

[4] Hartigan, J.A. (1975): *Clustering Algorithms.* New York: Wiley.

[5] Kogovšek, T., Ferligoj, A., Coenders, G., and Saris, W. (2002): Estimating the reliability and validity of personal support measures: Full information ML estimation with planned incomplete data. *Social Networks*, **24**, 1-20.

[6] Korenjak-Černe, S. and Batagelj, V. (1998): Clustering large datasets of mixed units. In A. Rizzi, M. Vichi, and H.H. Bock (Eds.): *Advances in Data Science and Classification.* Springer, 43-48.

[7] Marsden, P.V. and Campbell, K.E. (1984): Measuring tie strength. *Social Forces*, **63**, 482-501.

[8] Müller, C., Wellman, B., and Marin, A. (1999): How to use SPSS to study ego-centered networks. In *Bulletin de Méthodologies Sociologiques*, BMS, **64**, 63-76.