

Clustering of Attribute and/or Relational Data

Anuška Ferligoj¹

University of Ljubljana, e-mail: anuska.ferligoj@fdv.uni-lj.si

Abstract: A large class of clustering problems can be formulated as an optimizational problem in which the best clustering is searched among all feasible clustering according to a selected criterion function. This clustering approach can be applied to a variety of very interesting clustering problems, as it is possible to adapt it to a concrete clustering problem by an appropriate specification of the criterion function and/or by the definition of the set of feasible clusterings. Both, the blockmodeling problem (clustering of the relational data) and the clustering with relational constraint problem (clustering of the attribute and relational data) can be very successfully treated by this approach. It also opens many new developments in these areas.

Keywords: clustering problem, optimization, agglomerative approach, relocation approach, blockmodeling, relational constraint

1. Introduction

Grouping units into clusters so that those within a cluster are as similar to each other as possible, while units in different clusters as dissimilar as possible, is a very old problem. Although the clustering problem is intuitively simple and understandable, providing solution(s) remains a very exciting activity. The field of cluster analysis has its one society, the *International Federation of Classification Societies* which was formed in 1985 from several national classification societies. The society organizes every second year its conference and publishes two journals: the *Journal of Classification*, which was established in 1984 and the journal *Advances in Data Analysis and Classification* established in 2007. Clustering of relational data is one of the clustering topics that was mostly developed in the field of social network analysis. There, for the clustering of relational data the term blockmodeling is used. On the other side the clustering with relational constraint, where a solution is searched according to the attribute and the relational data, was mostly developed in the field of cluster analysis. A unified approach is presented.

2. Clustering Problem

Cluster analysis (known also as classification and taxonomy) deals mainly with the following general problem: given a set of units, $\mathcal{U}$, determine subsets, called clusters, C , which are homogeneous and/or well separated according to the measured variables. The set of clusters forms a clustering. This problem can be formulated as an optimization problem:

¹ Address of correspondence: Kardeljeva pl. 5, 1000 Ljubljana, Slovenia

Determine the clustering $\mathbf{C}^*$ for which

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

where $\mathbf{C}$ is a clustering of a given *set of units*, $\mathcal{U}$, Φ is the set of all feasible clusterings and $P : \Phi \rightarrow R$ is a *criterion function*.

There are several types of clusterings, e.g., partition, hierarchy, pyramid, fuzzy clustering, clustering with overlapping clusters. The most frequently used clusterings are partitions and hierarchies. A clustering $\mathbf{C} = \{C_1, C_2, \dots, C_k\}$ is a *partition* of the set of units $\mathcal{U}$ if $\bigcup_i C_i = \mathcal{U}$ and $i \neq j \Rightarrow C_i \cap C_j = \emptyset$. A clustering $\mathbf{H} = \{C_1, C_2, \dots, C_k\}$ is a *hierarchy* if for each pair of clusters C_i and C_j from $\mathbf{H}$ it holds $C_i \cap C_j \in \{C_i, C_j, \emptyset\}$ and it is a complete hierarchy if for each unit x it holds $\{x\} \in \mathbf{H}$, and $\mathcal{U} \in \mathbf{H}$.

Clustering criterion functions can be constructed *indirectly*, e.g., as a function of a suitable (dis)similarity measure between pairs of units (e.g., euclidean distance) or *directly*. In most cases, the criterion function is defined indirectly. For partitions into k clusters, the Ward criterion function (Ward, 1963)

$$P(\mathbf{C}) = \sum_{C \in \mathbf{C}} \sum_{x \in C} d(x, t_C)$$

is usually used, where t_C is the center of the cluster C and is defined as

$$t_C = (\bar{u}_{1C}, \bar{u}_{2C}, \dots, \bar{u}_{mC})$$

where $\bar{u}_{iC}$ is the average of the variable U_i , $i = 1, \dots, m$, for the units from the cluster C . d is the squared euclidean distance.

As the set of feasible clusterings is finite a solution of the clustering problem always exists. Since this set is usually very large it is not easy to find an optimal solution. In general, most of the clustering problems are NP-hard. For this reason, different efficient *heuristic* algorithms are used. There are many such algorithms and approaches. Among these, the agglomerative (hierarchical) and the relocation approach are most often used (see, e.g., Doreian et al., 2005).

The optimizational approach to clustering problem offers two possibilities to adapt to a concrete clustering problem: the definition of the criterion function P and the specification of the set of feasible clusterings Φ . The usual clustering problem seeks for a clustering according to the selected variables or attributes. Blockmodeling is searching for a clustering according to the relational data only. The solution can be obtained by an appropriately defined criterion function, described in Section 2. For clustering with relational constraint (attribute and relational data) an appropriately defined set of feasible clusterings is used (see Section 3).

3. Blockmodeling

Let $\mathcal{U}$ be a finite set of units and let the units be related by a binary relation $R \subseteq \mathcal{U} \times \mathcal{U}$ which determines a network $\mathbf{N} = (\mathcal{U}, R)$. One of the main procedural goals of social network analysis is to identify, in a given network, clusters of units that share structural characteristics defined in terms of the relation R . The units within a cluster have the same or similar connection patterns to the units of other clusters. A clustering $\mathbf{C}$ partitions also the relation R into *blocks* $R(C_i, C_j) = R \cap C_i \times C_j$. Each block is defined in terms of

units belonging to clusters C_i and C_j and consists of all arcs from units in cluster C_i to units in cluster C_j . If $i = j$, the block $R(C_i, C_i)$ is called a *diagonal* block.

A *blockmodel* consists of structures obtained by shrinking all units from the same cluster of the clustering $\mathbf{C}$. For an exact definition of a blockmodel we must be precise about which blocks produce an arc in the *reduced graph* and which do not. The reduced graph can be presented also by a relational matrix, called an *image matrix*.

The partition is constructed by using structural information contained in R only, and units in the same cluster are equivalent to each other in terms of R alone. These units share a common structural position within the network.

Blockmodeling, as a set of empirical procedures, is based on the idea that units in a network can be grouped according to the extent to which they are equivalent, in terms of some *meaningful* definition of equivalence. In general different definitions of equivalence usually lead to distinct partitions.

Lorrain and White (1971) provided a definition of *structural equivalence*: Units are equivalent if they are connected to the rest of the network in *identical* ways. x and y are structurally equivalent if and only if:

$$\begin{array}{ll} \text{s1. } xRy \Leftrightarrow yRx & \text{s3. } \forall z \in \mathcal{U} \setminus \{x, y\} : (xRz \Leftrightarrow yRz) \\ \text{s2. } xRx \Leftrightarrow yRy & \text{s4. } \forall z \in \mathcal{U} \setminus \{x, y\} : (zRx \Leftrightarrow zRy) \end{array}$$

From this definition it follows that only four possible ideal blocks can appear (Batagelj et al., 1992, Doreian et al., 2005)

$$\begin{array}{ll} \text{Type 0. } b_{ij} = 0 & \text{Type 2. } b_{ij} = 1 - \delta_{ij} \\ \text{Type 1. } b_{ij} = \delta_{ij} & \text{Type 3. } b_{ij} = 1 \end{array}$$

where δ_{ij} is the Kronecker delta function and $i, j \in C$. The blocks of types 0 and 1 are called the *null* blocks and the blocks of types 2 and 3 the *complete* blocks. For the nondiagonal blocks $R(C_u, C_v)$, $u \neq v$, only blocks of type 0 and type 3 are admissible.

Attempts to generalize the structural equivalence date back at least to Sailer (1978) and have taken various forms. Integral to all formulations is the idea that units are equivalent if they link in equivalent ways to other units that are also equivalent. Regular equivalence, as defined by White and Reitz (1983), is one such generalization.

The equivalence relation $\approx$ on $\mathcal{U}$ is a *regular equivalence* on network $\mathbf{N} = (\mathcal{U}, R)$ if and only if for all $x, y, z \in \mathcal{U}$, $x \approx y$ implies both

$$\text{R1. } xRz \Rightarrow \exists w \in \mathcal{U} : (yRw \wedge w \approx z) \quad \text{R2. } zRx \Rightarrow \exists w \in \mathcal{U} : (wRy \wedge w \approx z)$$

As was the case with structural equivalence, regular equivalence implies the existence of ideal blocks. The nature of these ideal blocks follows from the following theorem (Batagelj et al., 1992):

Theorem 1 *Let $\mathbf{C} = \{C_i\}$ be a partition corresponding to a regular equivalence $\approx$ on the network $\mathbf{N} = (\mathcal{U}, R)$. Then each block $R(C_u, C_v)$ is either null or it has the property that there is at least one 1 in each of its rows and in each of its columns. Conversely, if for a given clustering $\mathbf{C}$, each block has this property then the corresponding equivalence relation is a regular equivalence.*

Until now, a definition of equivalence was assumed for the *entire* network and the network was analyzed in terms of the permitted ideal blocks. Doreian, Batagelj and Ferligoj [9] generalized the idea of a blockmodel to one where the blocks can conform to more types

beyond the three mentioned above, and one where there is no single a priori definition of ‘equivalence’ for the entire network.

The problem of establishing a partition of units in a network, in terms of a considered equivalence, is a special case of the clustering problem – such that the criterion function reflects the considered equivalence. Such criterion functions can be constructed *indirectly* as a function of a compatible (dis)similarity measure between pairs of units. A dissimilarity d is *compatible* with the equivalence $\equiv$ if $x_i \equiv x_j \Leftrightarrow d(x_i, x_j) = 0$. Not many dissimilarities are compatible with the equivalences mentioned above. After calculating appropriate dissimilarities one of the clustering approaches (e.g., hierarchical or relocation approach) can be used to obtain a clustering solution.

The other possible way of constructing the criterion function is the direct way that directly reflects the considered equivalence. Such a criterion function measures the fit of a clustering to an ideal one with perfect relations within each cluster and between clusters according to the selected type of equivalence. The criterion function $P(\mathbf{C})$ defined should be sensitive to the considered equivalence:

$$P(\mathbf{C}) = 0 \Leftrightarrow \mathbf{C} \text{ determines the equivalence.}$$

Given a clustering $\mathbf{C} = \{C_1, C_2, \dots, C_k\}$, let $\mathcal{B}(C_u, C_v)$ denote the set of all ideal blocks corresponding to block $R(C_u, C_v)$. Then the global error of clustering $\mathbf{C}$ can be expressed as

$$P(\mathbf{C}) = \sum_{C_u, C_v \in \mathbf{C}} \min_{B \in \mathcal{B}(C_u, C_v)} d(R(C_u, C_v), B)$$

where the term $d(R(C_u, C_v), B)$ measures the difference (error or inconsistency) between the block $R(C_u, C_v)$ and the ideal block B .

In the case of the direct clustering approach, where an appropriate criterion function that captures the selected equivalence is constructed, relocation approach can be used to solve the given blockmodeling problem (Doreian et al., 2005).

The *inductive* approaches for establishing blockmodels for a set of social relations defined over a set of units were discussed above. Some form of equivalence is specified and clusterings are sought that are consistent with a specified equivalence. Another view of blockmodeling is *deductive* in the sense of starting with a blockmodel that is specified in terms of substance prior to an analysis. In this case given a network, set of types of ideal blocks, and a family of reduced models, a clustering can be determined which minimizes the criterion function. For details see (Batagelj et al., 1998, Doreian et al. 2005).

It is also possible to formulate a generalized blockmodeling problem where the network is defined by several sets of units and ties between them. Therefore, several partitions – for each set of units a partition has to be determined. The generalized blockmodeling approach was adapted for 2-way networks (Doreian et al., 2004), and only for structural equivalence and the indirect approach for 3-way networks (Batagelj et al., 2007).

Till now only binary networks were treated. Another interesting problem is the development of generalized blockmodeling of valued networks. Žiberna (2007) proposed several approaches to generalized blockmodeling of valued networks, where values of the ties are assumed to be measured on at least interval scale. The first approach is a straightforward generalization of the generalized blockmodeling of binary networks (Batagelj and Ferligoj, 2000) to valued blockmodeling. The second approach is homogeneity blockmodeling. The basic idea of homogeneity blockmodeling is that the inconsistency of an empirical block with its ideal block can be measured by within block variability of appropriate values.

4. Clustering with Relational Constraints

The constrained clustering problem can be expressed as clustering problem where the constraints are considered in the definition of the set of the feasible clusterings. In the case of clustering with the relational constraint, the problem is to find clusterings as similar as possible according to attribute data and also considering the ties from a relation R . The clustering with constraints problem seeks to determine the clustering $\mathbf{C}^*$ for which the criterion function P has the minimal value among all clusterings from the set of feasible clusterings $\mathbf{C} \in \Phi$, where Φ is determined by the relational constraints. Generally, such a set of feasible clusterings can be defined as:

$$\Phi(R) = \{\mathbf{C} : \mathbf{C} \text{ is a partition of } \mathcal{U} \text{ and each cluster } C \in \mathbf{C} \text{ is a subgraph } (C, R \cap C \times C) \text{ in the graph } (\mathcal{U}, R) \text{ with the required type of connectedness}\}$$

We can define different types of sets of feasible clusterings for the same relation R (Ferligoj and Batagelj, 1983). Some examples of clusterings with relational constraint $\Phi^i(R)$ are

type of clusterings	type of connectedness
$\Phi^1(R)$	weakly connected units
$\Phi^2(R)$	weakly connected units that contain at most one center
$\Phi^3(R)$	strongly connected units
$\Phi^4(R)$	clique
$\Phi^5(R)$	the existence of a trail containing all the units of the cluster

In the clustering type $\Phi^2(R)$ a center of a cluster C is the set of units $L \subseteq C$ iff the subgraph induced by L is strongly connected and $R(L) \cap (C \setminus L) = \emptyset$ where $R(L) = \{y : \exists x \in L : xRy\}$.

Standard clustering algorithms can be adapted for solving relational constrained clustering problems (e.g., the agglomerative hierarchical and the relocation approach) (Ferligoj and Batagelj, 1992, 1983).

Recently Batagelj, Ferligoj and Mrvar (2008) adapted the clustering with relational constraint approach for very large sets of units.

5. Software

Most of described clustering of attribute and relational data procedures are implemented in Pajek – program for analysis and visualization of large networks (Batagelj and Mrvar, 1998). It is freely available, for noncommercial use, at: <http://pajek.imfm.si>.

References

- Batagelj V., Doreian P. and Ferligoj A. (1992a) An optimizational approach to regular equivalence, *Social Networks*, 14, 121–135.
- Batagelj V. and Ferligoj A. (2000) Clustering relational data, in: *Data Analysis: Scientific Modeling and Practical Applications*, Gaul W., Opitz O. and Schader M., eds., Berlin: Springer, 3–15.

- Batagelj V., Ferligoj A. and Doreian P. (1992b) Direct and indirect methods for structural equivalence, *Social Networks*, 14, 63–90.
- Batagelj V., Ferligoj A. and Doreian P. (1998) Fitting to pre-specified blockmodels, in: *Classification and Related Methods*, Hayashi N., Jajima K., Bock H.H. and Baba Y., eds., Berlin: Springer, 199–206.
- Batagelj V., Ferligoj A. and Doreian P. (2007) Indirect blockmodeling of 3-way networks, in: *Selected Contributions in Data Analysis and Classification*, Brito P., Bertrand P., Cucumel G. and de Carvalho F., eds., Berlin: Springer, 151–159.
- Batagelj V., Ferligoj A. and Mrvar A. (2008) Hierarchical clustering in large networks, in: *XXVIII. SUNBELT - International Network for Social Network Analysis*, Hayashi N., Jajima K., Bock H.H. and Baba Y., eds., Proceedings, 159.
- Batagelj V. and Mrvar A. (1998) Pajek - a program for large network analysis, *Connections*, 21, 47–57.
- Doreian P., Batagelj V. and Ferligoj A. (2004) Generalized blockmodeling of two-mode network data, *Social Networks*, 26, 29–53.
- Doreian P., Batagelj V. and Ferligoj A. (2005) *Generalized Blockmodeling*, Cambridge: Cambridge University Press.
- Ferligoj A. and Batagelj V. (1982) Clustering with relational constraint, *Psychometrika*, 47, 413–426.
- Ferligoj A. and Batagelj V. (1983) Some types of clustering with relational constraints, *Psychometrika*, 48, 541–552.
- Lorrain F. and White H. (1971) Structural equivalence of individuals in social networks, *Journal of Mathematical Sociology*, 1, 49–80.
- Sailer L. (1978) Structural equivalence: Meaning and definition, computation and application, *Social Networks*, 1, 73–90.
- Žiberna A. (2007) Generalized blockmodeling of valued networks, *Social Networks*, 29, 105–126.
- Ward J. (1963) Hierarchical grouping to optimize an objective function, *Journal of American Statistical Association*, 58, 236–244.
- White D. and Reitz K. (1983) Graph and semigroup homomorphisms on networks of relations, *Social Networks*, 5, 193–234.