Multiple dissimilarity SOM for clustering and visualizing graphs with node and edge attributes

Nathalie Villa-Vialaneix NATHALIE.VILLA@TOULOUSE.INRA.FR
INRA, UR 0875 MIAT, BP 52627, 31326 Castanet Tolosan, FRANCE

Madalina Olteanu MADALINA.OLTEANU@UNIV-PARIS1.FR
SAMM, EA4543, Université Paris 1, 90 route de Tolbiac, 75013 Paris cedex 13, FRANCE

Introduction When wanting to understand the way a graph $\mathcal{G}$ is structured and how the relations it models organize groups of entities, clustering and visualization can be combined to provide the user with a global overview of the graph, on the form of a projected graph: a simplified graph is visualized in which the nodes correspond to a cluster of nodes in the original graph $\mathcal{G}$ (with a size proportional to the number of nodes that are classified inside this cluster) and the edges between two nodes have a width proportional to the number of links between the nodes of $\mathcal{G}$ classified in the two corresponding clusters. This approach can be trickier when additional attributes (numerical or factors) describe the nodes of $\mathcal{G}$ or when the edges of $\mathcal{G}$ are of different types and should be treated separately: the simplified representation should then represent similarities for all sets of information. In this proposal, we present a variant of Self-Organizing Maps (SOM), which is adapted to data described by one or several (dis)similarities or kernels recently published in (Olteanu & Villa-Vialaneix, 2015) and which is able to combine clustering and visualization for this kind of graphs.

SOM for dissimilarity data The relational/kernel SOMs are two adaptations of the well known SOM algorithm (Kohonen, 2001) to data that are described by a dissimilarity matrix $(\delta(x_i, x_j))_{i,j=1,\dots,n}$ (or a kernel) (Hammer & Hasenfuss, 2010) among others. It aims at projecting the data $(x_i)_i$ into a two-dimensional grid made of U units and equipped with a distance that defines a topology on the grid. Each unit is represented by a prototype p_u which can be interpreted as a generalized centroid of the observations in the Euclidean case. When the data $(x_i)_i$ are not defined in a Euclidean space, (Goldfarb, 1984) (dissimilarity case) justifies the definition of p_u as a convex combination of some transformation of the data $\sum_{i=1}^{n} \alpha_i \phi(x_i)$ with $\alpha_i \geq 0$ and $\sum_i \alpha_i = 1$. The algorithm then iterates over an *affectation step* which affects one data picked at random to the unit with the closest prototype (in the kernel/dissimilarity sense) and a *representation step* which mimics a stochastic gradient descent in the feature space defined by ϕ. This method is implemented in the R package **SOMbrero**. It can be used to cluster the vertices of the graph on a SOM grid, using any dissimilarity (shortest path lengths or distance induced by a kernel for graphs). Functions that allow to display the projected graph using the *a priori* positions of the clusters on the grid are included in this package.

Graphs with attributes When additional information is provided on the nodes or when the edges are of different types, the graph can be described by several dissimilarity matrices $(D^k)_{k=1,\dots,K}$, each describing the dissemblance between nodes for a given feature (each kind of edges or each attribute describing the nodes). We have proposed to define a new dissimilarity based on the convex definition of the different dissimilarities $\sum_{k=1}^{K} \beta_k D^k$ and to optimize the convex combination $(\beta_k)_k$ including a stochastic gradient descent step in the algorithm, performed after the representation step (see (Rakotomamonjy et al., 2008) for a similar idea). The approach has been proven successful to cluster nodes of a graph with numeric and factor descriptors on a synthetic graph. It is still to be tested for graphs with multiple type edges.

References

Goldfarb, L. A unified approach to pattern recognition. *Pattern Recognition*, 17(5):575–582, 1984. doi: 10.1016/0031-3203(84)90056-6.

Hammer, B. and Hasenfuss, A. Topographic mapping of large dissimilarity data sets. *Neural Computation*, 22(9):2229–2284, September 2010.

Kohonen, T. *Self-Organizing Maps, 3rd Edition*, volume 30. Springer, Berlin, Heidelberg, New York, 2001.

Olteanu, M. and Villa-Vialaneix, N. On-line relational and multiple relational SOM. *Neurocomputing*, 147:15–30, 2015. doi: 10.1016/j.neucom.2013.11.047.

Rakotomamonjy, A., Bach, F.R., Canu, S., and Grandvalet, Y. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.