IMPROVING CLUSTERING IN TEMPORAL NETWORKS THROUGH AN EVOLUTIONARY ALGORITHM

Luca Brusa[1] and Fulvia Pennoni[1]

[1] Department of Statistics and Quantitative Methods, University of Milano-Biccoca
(e-mail: luca.brusa@unimib.it, fulvia.pennoni@unimib.it)

ABSTRACT: The dynamic stochastic blockmodel is commonly used to analyze longitudinal network data when multiple snapshots are observed over time. The variational expectation-maximization (VEM) algorithm is typically employed for maximum likelihood inference to allocate nodes to groups dynamically. To address the problem of multiple local maxima, which may arise in this context, we propose modifying the VEM according to an evolutionary algorithm to explore the whole parameter space. A simulation study on dynamic networks and an application illustrate the proposal comparing the performance with that of the VEM algorithm.

KEYWORDS: local maxima, longitudinal networks, node classification, stochastic blockmodel, variational expectation-maximization algorithm.

1 Introduction

The dynamic stochastic blockmodel (Matias & Miele, 2017) extends the stochastic blockmodel (SB, Nowicki & Snijders, 2001) for the analysis of longitudinal network data when multiple snapshots are observed over time. This model aims to identify homogeneous blocks of nodes and to analyze interactions between nodes and their evolution. At each time occasion, nodes are partitioned into a set of groups whose number is estimated; the probability of observing an edge between a couple of nodes depends on the assigned groups.

In the inferential context, the variational expectation-maximization (VEM, Jordan *et al.*, 1999) algorithm has been proposed for maximum likelihood estimation. However, a drawback of this method is that it can be trapped in one of the multiple local maxima. To account for this problem we propose a modified version of the VEM through an evolutionary algorithm (EA, Ashlock, 2004). We perform a Monte Carlo simulation study to evaluate the performance of the proposed evolutionary VEM (EVEM) algorithm in avoiding local maxima and improving the accuracy of the posterior classification. We also show an application estimating the dynamic SB with data related to face-to-face contacts between employees to investigate transmission of an infectious disease.

2 Notation and inference in dynamic stochastic blockmodel

Considering n nodes observed at T discrete times, let $\mathbf{Y}$ denote an adjacency array of dimensions $n \times n \times T$, where $\mathbf{Y}^{(t)}$ is the adjacency matrix at time t and $Y_{ij}^{(t)} = 1$ if there is an edge between nodes i and j (symmetric association) at time t and $Y_{ij}^{(t)} = 0$ otherwise $(i, j = 1, \dots, n,\ i \neq j)$. The dynamic SB assumes that block membership depends on a set of independent and identically distributed discrete latent variables $Z_i^{(t)}$ following a Markov chain with k support points. In this way, each node is partitioned into one of k latent blocks at every time occasion according to the initial and the transition probabilities denoted as α_u and π_{uv}, $u, v = 1, \dots, k$, respectively. Under the local independence assumption and conditionally on the latent blocks to which nodes i and j belong at time t, the variables $Y_{ij}^{(t)}$ are assumed to be independent and Bernoulli distributed with connection probabilities denoted as β_{uv}.

For maximum likelihood inference of SB the VEM was proposed in Matias & Miele, 2017 to maximize a lower bound of the log-likelihood function denoted as $\mathcal{J}(\theta)$, where θ collects the model parameters. More recently, Bartolucci & Pandolfi, 2020, proposed an exact formulation of the VEM algorithm to improve clustering units across time occasions. They initialize the starting values for the model parameters through the k-means method since random initialization is usually ineffective in this context. However, this approach does not prevent the VEM algorithm from being trapped in the local maxima that frequently arise with complex data structures.

3 Proposed evolutionary VEM algorithm

The proposed EVEM algorithm is defined by the following features: (*i*) an initial "population" denoted as P_0 of N candidate solutions for the maximization problem at issue, here specified as possible arrays of cluster memberships; (*ii*) a mutation operator that introduces variations to the existing candidates and generates new solutions by randomly selecting an observation and providing an updated cluster membership; (*iii*) selection of the best solutions based on a quality measure that favors candidates with higher values of $\mathcal{J}(\theta)$.

In order to explore the whole parameter space the first candidate for population P_0 is obtained according to the k-means deterministic initialization; in particular, the adjacency matrices $\mathbf{Y}^{(t)}$ for $t = 1, \dots, T$ are row-concatenated together, and the k-means algorithm is applied on the rows of the resulting $nT \times n$ matrix. Then, the remaining $N - 1$ candidates are obtained through mutation. The procedure alternates the following steps until convergence:

1. $P_1 \leftarrow$ **Update**(P_0): perform a small number of iterations of the VEM algorithm on each individual of population P_0.
2. $P_2 \leftarrow$ **Mutate**(P_1): add variation in each individual of population P_1 to encourage a broader exploration of the parameter space.
3. $P_3 \leftarrow$ **Update**(P_2): perform a small number of iterations of the VEM algorithm on each individual of population P_2.
4. $P_4 \leftarrow$ **Select**$(P_1 \cup P_3)$: consider individuals of both populations P_1 and P_3, and retain the N showing the highest value of $\mathcal{J}(\theta)$ for the next generation.

Convergence is assessed considering the best solution of population P_4, analyzing the relative difference of $\mathcal{J}(\theta)$ at two consecutive steps and that between the corresponding parameter vectors.

4 Simulation study and application

In analogy with the design used in Bartolucci & Pandolfi, 2020, a Monte Carlo simulation study is conducted, varying the number of nodes ($n = 20, 50$), the number of latent blocks ($k = 2, 3$), the block persistence (high or low), and the connectivity parameters (intra-group greater or smaller than inter-group). For each of the 16 resulting scenarios, we randomly draw 50 networks and estimate the dynamic SB with both the VEM and the EVEM algorithms. The effectiveness of the proposed approach is evaluated in terms of the Adjusted Rand Index (ARI, Hubert & Arabie, 1985) between the true and the estimated classification at each time occasion.

Simulation results show that the EVEM algorithm outperforms the existing VEM algorithm in most scenarios, especially those with higher complexity. For example, considering a scenario characterized by 50 nodes, 3 latent blocks, low persistence of latent states, and higher intra-group than inter-groups connection probabilities, the ARI equals 0.688 using the VEM algorithm and 0.761 with the EVEM algorithm. In another scenario, with the same features but opposite connectivity parameter setting, ARI is 0.707 with VEM and 0.784 with the EVEM. In both cases, the improvements are statistically significant. When using the EVEM algorithm, we also observe a decrease of the mean squared error between the estimated and true model parameters, computed as an aggregated measure over all the model parameters.

Real data refer to face-to-face contacts between $n = 90$ employees in a building of the *Institut de veille sanitaire* (French Institute for Public Health

Surveillance) for ten working days ($T = 10$), from June 24 to July 3, 2013 (data are available at the website: `http://www.sociopatterns.org/datasets/contacts-in-a-workplace/`). The building hosts three scientific departments ("DISQ", "DMCT", and "DES"), logistics ("SFLE") and human resources ("SRH"). The adjacency array is built by setting each element $Y_{ij}^{(t)}$ equal to 1 if at least one face-to-face contact was registered between employees i and j at time t, and 0 otherwise.

A dynamic SB with 5 latent blocks is estimated using both VEM and EVEM algorithms. The resulting classification of employees helps understand how a certain infectious disease may spread across different departments of the same building. We observe that the value of $\mathcal{J}(\hat{\theta})$ at convergence increases from -2613 to -2600 when the EVEM algorithm is employed. This is reflected in a more accurate classification of the employees in each group of the network. The EVEM algorithm identifies a specific latent block for employees from the "DISQ" department, while the VEM algorithm allocates them with employees from the "DMCT" department. Additionally, the EVEM algorithm correctly assigns all employees from the "DSE" department to a single latent block, whereas the VEM algorithm splits them into two distinct blocks.

References

ASHLOCK, D. 2004. *Evolutionary Computation for Modeling and Optimization*. Springer, New York.

BARTOLUCCI, F., & PANDOLFI, S. 2020. An exact algorithm for time-dependent variational inference for the dynamic stochastic block model. *Pattern Recognit. Lett.*, **138**, 362–369.

HUBERT, L., & ARABIE, P. 1985. Comparing partitions. *J. Classif.*, **2**, 193–218.

JORDAN, M.I., GHAHRAMANI, Z., JAAKKOLA, T.S., & SAUL, L.K. 1999. An introduction to variational methods for graphical models. *Mach. Learn.*, **37**, 183–233.

MATIAS, C., & MIELE, V. 2017. Statistical clustering of temporal networks through a dynamic stochastic block model. *J. R. Stat. Soc. Series B*, **79**, 1119–1141.

NOWICKI, K., & SNIJDERS, T.A.B. 2001. Estimation and prediction for stochastic blockstructures. *J. Am. Stat. Assoc.*, **96**, 1077–1087.