ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Mixture of stochastic block models for multiview clustering

Kylliann De Santiago[1], Marie Szafranski[2,1] and Christophe Ambroise[1]

1- Université Paris-Saclay, CNRS, Univ Évry,
Laboratoire de Mathématiques et Modélisation d'Évry,
91037, Évry-Courcouronnes, France.

2- ENSIIE, 91025, Évry-Courcouronnes, France.

Abstract. In this work, we propose an original method for aggregating multiple clustering coming from different sources of information. Each partition is encoded by a co-membership matrix between observations. Our approach uses a mixture of Stochastic Block Models (SBM) to group co-membership matrices with similar information into components and to partition observations into different clusters, taking into account their specificities within the components. The parameters are estimated using a Variational Bayesian EM algorithm. The Bayesian framework allows for selecting an optimal numbers of clusters and components.

1 Multiview data and late fusion clustering

Most everyday learning situations are achieved by integrating different sources of information, such as vision, touch and hearing. A source of information in a given format will be referred to as a modality or a *view*. Multimodal or multiview machine learning aims to learn models from multiple views (e.g. text, sound, image, etc.) in order to represent, translate, align, fusion, or co-learn [1, 2].

The corresponding learning models vary based on their fusion strategy. The three main methods are early, intermediate, and late fusion of views. Late fusion is well suited to clustering since each view is often associated to dedicated efficient clustering algorithms. In this work, we are interested to build a coordinated representation of the resulting partitions through a probabilistic model. This coordinated representation should provide details about consensus and complementary information existing between the different views both at a global and local level [2].

Several methods have been developed to discover relationships across multiple clustering results. *Consensus clustering* is a family of methods which are based on the construction of a consensus matrix that represents the degree of agreement among different clustering algorithms, as in [3]. *Tensor-based meta-clustering* utilizes multilinear algebra decomposition techniques to identify patterns and relationships across each layer of this tensor ; in case of meta-clustering, all layers are often adjacency matrices, as in [4]. Finally, *Mixture Multilayer Stochastic Block Models* expands the traditional SBM by considering multilayer networks, allowing the layers to come from a mixture model, as in [5].

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

We propose a Bayesian multilayer SBM approach, MIxture of Multiview Integrator SBM (mimi-SBM), that takes into account several sources of information, and where the membership clustering is traversing, as illustrated in Figure 1.

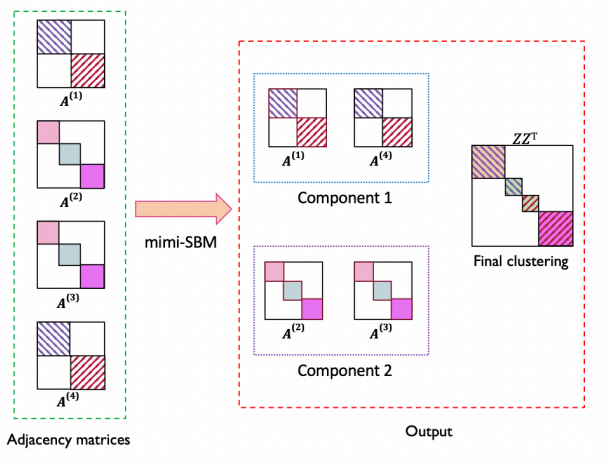


Fig. 1: Illustration of mimi-SBM. Left: Four adjacency matrices coming from four different views which can be organized into two components. Right: identification of the two components from the views (local and complementary information) and clustering of the observations (global and consensus information).

The next section introduces notations and model, followed by synthetic experiments for comparing and evaluating the performances of the proposed approach. We conclude with a discussion section.

2 A mixture of stochastic block models

Performing clustering with mixture models amounts to an inference problem since the cluster labels are the model latent variables. In our context, we consider two sets of latent variables respectively corresponding to the structure of the observations and the structure of the views. We assume that the views are generated by a mixture model, in which each component is itself a stochastic block model. This allows us to capture the complex dependencies between the views, and to accurately model the underlying structure of the data.

Let $\mathbf{A} \in \{0,1\}^{N \times N \times V}$ be a tensor where N is the number of vertices (observations), and V the number of views. $\mathbf{A}$ is defined as a natural extension of an adjacency matrix for multiple graphs $(\mathcal{G}^1, \cdots, \mathcal{G}^V)$ with corresponding vertices

$$A_{ijv} = \begin{cases} 1, & \text{if individuals } i, j \text{ are linked in } \mathcal{G}^v, \\ 0, & \text{otherwise.} \end{cases}$$

The nodes and their corresponding order are the same throughout all views.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Let denote $\mathbf{W} \in \{0,1\}^{V \times Q}$, the indicator membership matrix for the views, and $\mathbf{Z} \in \{0,1\}^{N \times K}$, the indicator membership matrix of observations, where K is the number of view traversing clusters and Q the number of components of the view mixture. Each line of matrix $\mathbf{W}$ follows a multinomial distribution, $\mathbf{W}_v \sim \mathcal{M}(1, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_Q))$.

Although we use multiple views with their own cluster structure, we assume a traversing structure for the latent variables across all views. By leveraging all available sources of information, we aim to achieve a more comprehensive understanding of the data and obtain community structures that are consistent across all views. The individuals are thus assumed to come from a number K of sub-populations. Each latent class vector follows a multinomial distribution, $\mathbf{Z}_i \sim \mathcal{M}(1, \boldsymbol{\pi} = (\pi_1, \ldots, \pi_K))$.

Eventually, the probability of an edge between individuals i and j on the view v, given the latent variables, is

$$\mathbb{P}(\mathbf{A} \mid \mathbf{Z}, \mathbf{W}, \boldsymbol{\Theta}) = \prod_{\substack{i=1,\\ i<j}}^{N} \prod_{k,l=1}^{K} \prod_{v=1}^{V} \prod_{s=1}^{Q} \left(\alpha_{kls}^{A_{ijv}} \left(1 - \alpha_{kls}\right)^{1 - A_{ijv}} \right)^{Z_{ik} Z_{jl} W_{vs}} .$$

The proposed Bayesian model considers the following classical distributions in the context of SBM [6]:

$$\mathbb{P}(\boldsymbol{\pi} \mid \boldsymbol{\beta}^0 = (\beta_1^0, \ldots, \beta_K^0)) = \mathrm{Dir}(\boldsymbol{\pi}; \boldsymbol{\beta}^0) \,,$$
$$\mathbb{P}(\boldsymbol{\rho} \mid \boldsymbol{\theta}^0 = (\theta_1^0, \ldots, \theta_Q^0)) = \mathrm{Dir}(\boldsymbol{\rho}; \boldsymbol{\theta}^0) \,,$$

where $\mathrm{Dir}(\cdot)$ define the Dirichlet distribution which is the conjugate prior for the multinomial distribution. Concerning the parameters of the SBMs, we use independent Beta priors to model the connectivity matrices

$$\mathbb{P}(\boldsymbol{\alpha} \mid \boldsymbol{\eta}^0 = (\eta_{kls}^0), \boldsymbol{\xi}^0 = (\xi_{kls}^0)) = \prod_{k,k<l} \prod_{s} \mathrm{Beta}(\alpha_{kls}; \eta_{kls}^0, \xi_{kls}^0) \,.$$

Jeffrey's prior seeks to establish a non-biased prior distribution for a parameter, thus reducing subjectivity in analysis [7].

In case of a Dirichlet distribution, we have $\beta_k^0 = \theta_s^0 = 1/2 \,, \forall k, s$ and for Beta distribution we can choose $\eta_{kls}^0 = \xi_{kls}^0 = 1/2 \,, \forall k, l, s$ too [6].

The marginal likelihood of observed data is defined as

$$\mathbb{P}\left(\mathbf{A}\right) = \sum_{\mathbf{Z}} \sum_{\mathbf{W}} \int \int \int \mathbb{P}\left(\mathbf{A}, \mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\right) \, d\boldsymbol{\alpha} \, d\boldsymbol{\pi} \, d\boldsymbol{\rho} \,.$$

Computing the marginal likelihood is a challenging problem in SBM. Integrals in the formula for the marginal likelihood are difficult or impossible to compute analytically, and sums over $\mathbf{Z}$ and $\mathbf{W}$ are often intractable as soon as the number of parameters or observations is large.

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

Given a variational distribution $q(\cdot)$ over $\{\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}\}$, we can decompose the marginal log-likelihood into Evidence Lower BOund (ELBO) and KL-divergence between variational and posterior distribution:

$$\log\left(\mathbb{P}(\mathbf{A})\right) = \mathcal{L}\left(q(\cdot)\right) + \mathbf{KL}\left(q(\cdot) \| \, \mathbb{P}(\cdot \mid \mathbf{A})\right) , \tag{1}$$

A convenient variational distribution, often from the exponential family, is chosen. Its parameters are optimized to minimize KL-divergence.

Using the mean-field approximation, we assume that $q(\cdot)$ can be factorised as

$$q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\alpha}, \boldsymbol{\pi}, \boldsymbol{\rho}) = \prod_{i=1}^{N} q(\mathbf{Z}_i) \prod_{v=1}^{V} q(\mathbf{W}_v) \prod_{s=1}^{Q} \prod_{k,k\leq l}^{K} q(\alpha_{kls}) \; q(\boldsymbol{\pi}) \; q(\boldsymbol{\rho}) \, . \tag{2}$$

According to (1) and (2), given a well defined distribution $q(\cdot)$, the ELBO is

$$\begin{aligned} \mathcal{L}\left(q(\cdot)\right) = \; & \mathrm{lBeta}(\boldsymbol{\beta}) - \mathrm{lBeta}(\boldsymbol{\beta}^0) + \mathrm{lBeta}(\boldsymbol{\theta}) - \mathrm{lBeta}(\boldsymbol{\theta}^0) + \\ & \sum_{k\leq l}^{K} \sum_{s=1}^{Q} \log\left(\frac{\Gamma\left(\eta^0_{kls} + \xi^0_{kls}\right) \Gamma\left(\eta_{kls}\right) \Gamma\left(\xi_{kls}\right)}{\Gamma\left(\eta_{kls} + \xi_{kls}\right) \Gamma\left(\eta^0_{kls}\right) \Gamma\left(\xi^0_{kls}\right)}\right) - \\ & \sum_{i}^{N} \sum_{k}^{K} \tau_{ik} \log(\tau_{ik}) \; - \sum_{v}^{V} \sum_{s}^{Q} \nu_{vs} \log(\nu_{vs}) \, , \end{aligned}$$

where $\Gamma(\cdot)$ is the Gamma function, $\mathrm{lBeta}(\cdot)$ the log multinomial Beta function, and the estimated probabilities τ_{ik} and ν_{vs} are respectively related to the memberships of observation i to cluster k and view v to component s. This function is also called ILvb and can be used for model selection.

3 Simulation study

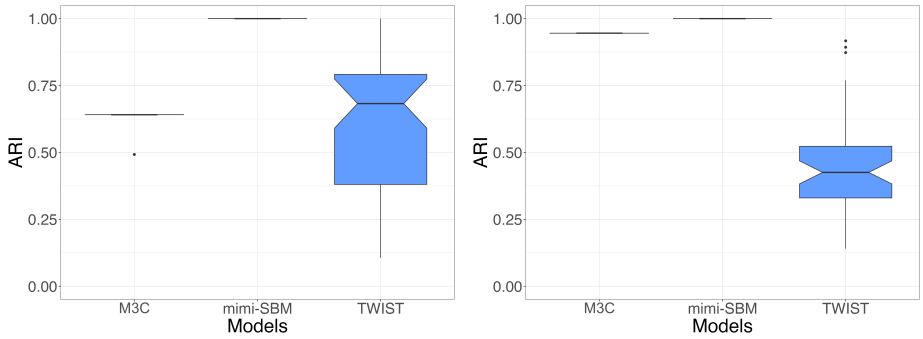
The Adjusted Rand Index (ARI) measures partition similarity in upcoming simulation. It quantifies agreement between model predictions and true partition, reflecting paired observations' grouping or separation. ARI near 1 signifies higher partition similarity.

For model comparison, artificial data is generated (adjacent matrices)from source mixtures and individual clustering. Various parameter values test different scenarios: N (observations), V (views), K (clusters), and Q (sources) are varied. For the other parameters, they have been set as follows : $\boldsymbol{\pi}, \boldsymbol{\rho}$ corresponding to an equiprobability of belonging to a cluster or component, thus $\{\pi_k\}_{k=1}^{K} = 1/K$ and $\{\rho_s\}_{s=1}^{Q} = 1/Q$.

Also, $\boldsymbol{\alpha}$ is a tensor composed of matrices. To have a simulation model that closely resembles a co-membership matrix, the matrices have 0.99 on the diagonal 0.01 and elsewhere. This setting strongly encourages links between individuals within the same cluster while allowing for some noise. The code for the simulations is available on GitHub in the repository *mimiSBM*. [1]

[1] https://github.com/Kdesantiago/mimiSBM.

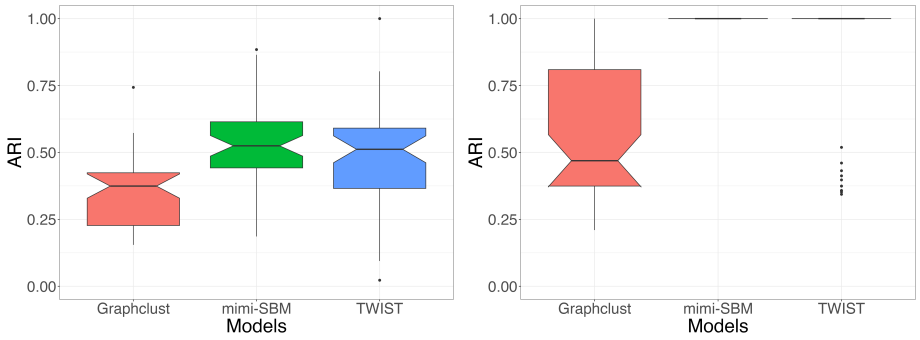
ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.



(a) $N = 200, V = 15, K = 5, Q = 3$ (b) $N = 200, V = 50, K = 10, Q = 10$

Fig. 2: Boxplot of ARI measure between true individual partition and output partition of M3C [3], mimi-SBM and TWIST [4] models.

In Figure 2 it has been observed that mini-SBM achieved the best clustering results for each considered experimental configuration. Indeed, mini-SBM recorded the highest ARI score for all data sizes, number of clusters and sources. M3C model improved with increased factors, while TWIST struggled with complex clustering, hinting at its limited suitability for difficult problems.



(a) $N = 50, V = 15, K = 5, Q = 3$ (b) $N = 200, V = 15, K = 5, Q = 3$

Fig. 3: Boxplot of ARI measure between true view clustering and output clustering of Graphclust [5], mimi-SBM and TWIST [4] models.

The second problem is the clustering of views. In Figure 3, Limited observations challenge source identification for all models. Graphclust stands out significantly, with non-overlapping boxplots compared to the other two models. As observations increase in the second simulation, model performances improve. Graphclust identifies true sources less than mimi-SBM and TWIST. TWIST mostly identifies true members but with occasional errors, seen as boxplot out-

ESANN 2023 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 4-6 October 2023, i6doc.com publ., ISBN 978-2-87587-088-9. Available from http://www.i6doc.com/en/.

liers.

The mimi-SBM model has demonstrated a clear ability in recovering the stratification of individuals and the components of the mixture of views. However, as with any statistical model, its performance, particularly on the mixture of views, improves with larger numbers of observations. Given the importance of accurately modeling mixture components in a variety of applications, these results highlight the potential utility of mimi-SBM in a range of contexts.

4 Conclusive remarks

In our simulation setting, the proposed Bayesian mixture multiview SBM approach has been shown to outperform methods based on tensor decomposition, hierarchical model-based SBM, and reference model in consensus clustering.

An interesting follow-up would be to extend this approach to the context of deep learning, specifically in the context of variational auto-encoders using the Bayesian formulation. Additionally, further research is needed to develop theoretical proofs regarding the convergence of parameters for the component-connection probability tensor model and explore the identifiability of this model.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.

[2] J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

[3] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52:91–118, 2003.

[4] B.-Y. Jing, T. Li, Z. Lyu, and D. Xia. Community detection on mixture multilayer networks via regularized tensor decomposition. *The Annals of Statistics*, 49(6):3181–3205, 2021.

[5] T. Rebafka. Model-based clustering of multiple networks with a hierarchical algorithm. 2023.

[6] P. Latouche, É. Birmele, and C. Ambroise. Variational bayesian inference and complexity control for stochastic block models. *Statistical Modelling*, 12(1):93–115, 2012.

[7] Harold Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 186(1007):453–461, 1946.