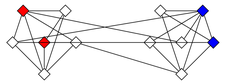
Learning the right layers: a data-driven layer-aggregation strategy for semi-supervised learning on multilayer graphs

Sara Venturini [1] Andrea Cristofari[2] Francesco Rinaldi[1] Francesco Tudisco[3]

[1]University of Padova, Italy [2]University of Rome "Tor Vergata", Italy [3]Gran Sasso Science Institute, Italy

Graph Semi-supervised Learning Problem



- $G = (V, E)$ graph
- $C = \{C_1, \dots, C_m\}$ set of classes of G
- set of input known labels for class : $Y_{ij} = 1$ if $i \in C_j$, and $Y_{ij} = 0$ otherwise

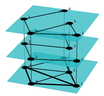
Aim: label the remaining vertices.

Optimization problem

$$\min_x \varphi(X) := \|X - Y\|_F^2 + \frac{\lambda}{2} \sum_{i,j} A^G_{ij} \|X_{:i} - X_{:j}\|^2$$

- Y columns one-hot encoder vectors each class
- A^G adjacency matrix of the graph
- $\lambda \geq 0$ regularization parameter

Extension to Multilayer Networks



$A = \{A^{G^1}, \dots, A^{G^K}\}$ with A^{G^k} adjacency matrix layer G^k

Generalized mean adjacency model

The generalized mean adjacency matrix

$$A(\alpha, \boldsymbol{\beta})_{ij} = \left(\sum_{k=1}^{K} \boldsymbol{\beta}_k \left(A^{(k)}_{ij}\right)^{\alpha} \right)^{1/\alpha} \quad \text{with } \alpha \in \mathbb{R}, \boldsymbol{\beta} \geq 0, e^T\boldsymbol{\beta} = 1$$

$\alpha \to -\infty$	$\alpha = -1, \boldsymbol{\beta}_k = 1/K$	$\alpha \to 0, \boldsymbol{\beta}_k = 1/K$	$\alpha = 1, \boldsymbol{\beta}_k = 1/K$	$\alpha \to +\infty$
$\min_{k=1,\dots,K} A^{(k)}_{ij}$	$\left(\frac{1}{K}\sum_{k=1}^{K} \frac{1}{A^{(k)}_{ij}}\right)^{-1}$	$\left(\prod_{k=1}^{K} A^{(k)}_{ij}\right)^{1/K}$	$\frac{1}{K}\sum_{k=1}^{K} A^{(k)}_{ij}$	$\max_{k=1,\dots,K} A^{(k)}_{ij}$
Minimum	Harmonic	Geometric	Arithmetic	Maximum

Bilevel optimization model

In order to learn the parameters $\boldsymbol{\theta} := (\alpha, \boldsymbol{\beta}, \lambda)$, we split the available input labels into training and test sets: Y^{tr} and Y^{te}, and consider the bilevel optimization model

$$\begin{aligned} \min_{\boldsymbol{\theta}} \quad & H(Y^{te}, X_{Y^{tr},\boldsymbol{\theta}}) \\ \text{s.t.} \quad & X_{Y^{tr},\boldsymbol{\theta}} = \operatorname{argmin}_X \varphi(X, Y^{tr}, \boldsymbol{\theta}) \\ & \boldsymbol{\theta} = (\alpha, \boldsymbol{\beta}, \lambda),\ \alpha \in \mathbb{R},\ \boldsymbol{\beta} \geq 0,\ \textstyle\sum_k \boldsymbol{\beta}_k = 1\ \lambda \in \mathbb{R} \end{aligned}$$

with

- H cross-entropy loss function
- $\varphi(X, Y, \theta) = \|X - Y\|_F^2 + \frac{\lambda}{2} \sum_{i,j} A(\alpha, \beta)_{ij} \|X_{:i} - X_{:j}\|^2$

BINOM: binomial cross-entropy loss for each community

$$h(y, x) = -\frac{1}{N} \sum_{i=1}^{N} \left(y_i \log(x_i) + (1 - y_i) \log(1 - x_i) \right).$$

MULTI: multiclass cross-entropy loss

$$H(Y, X) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} Y_{ij} \log\left(\frac{X_{ij}}{\sum_{j=1}^{N} X_{ij}} \right).$$

Lower level problem

The lower level problem

$$min_X \varphi(X, Y^{tr}, \boldsymbol{\theta})$$

is solved explicitly using Label Propagation,
over the graph induced by the generalized mean adjacency matrix $A(\alpha, \boldsymbol{\beta})$.

Upper level problem

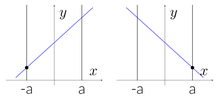
Feasible region:

$$S = \begin{cases} \alpha \in [-a, a] \\ \boldsymbol{\beta} \geq 0, e^T\boldsymbol{\beta} = 1 \\ \lambda \in [l_0, l_1] \end{cases}$$

We solve it using the Frank Wolfe algorithm with inexact gradient.
In each iteration we solve the linearized problem:

$$\hat{\theta} = \min_{\theta \in S} \widetilde{\nabla} H(\theta_n)^T (\theta - \theta_n)$$

which can be solved separately in the the variables $\theta = (\alpha, \boldsymbol{\beta}, \lambda)$.

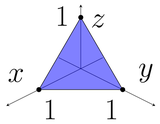


For variable $\alpha \in [-a, a]$, we need to solve

$$\begin{aligned} \min_{\alpha} \quad & \widetilde{\nabla}_\alpha H(\alpha_n, \boldsymbol{\beta}_n, \lambda_n)(\alpha - \alpha_n) \\ \text{s.t.} \quad & \alpha \in [-a, a] \end{aligned}$$

$$\hat{\alpha}_n = \begin{cases} -a & \text{if } \widetilde{\nabla}_\alpha H(\alpha_n, \boldsymbol{\beta}_n, \lambda_n) > 0 \\ a & \text{otherwise} \end{cases}$$

Same for $\lambda \in [l_0, l_1]$.



For the set of variables $\boldsymbol{\beta} \in \mathbb{R}^K$, we need to solve

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \widetilde{\nabla}_{\boldsymbol{\beta}} H(\alpha_n, \boldsymbol{\beta}_n, \lambda_n)^T (\boldsymbol{\beta} - \boldsymbol{\beta}_n) \\ \text{s.t.} \quad & \boldsymbol{\beta} \geq 0 \\ & e^T\boldsymbol{\beta} = 1 \end{aligned}$$

$$\hat{\boldsymbol{\beta}}_n = e_{\hat{j}}$$

where $\hat{j} = \operatorname{argmin}_{j=1,\dots,K} [\widetilde{\nabla}_{\boldsymbol{\beta}} H(\alpha_n, \boldsymbol{\beta}_n, \lambda_n)]_j$ and e_j vector of the canonical basis of $\mathbb{R}^K$.

Convergence Analysis

Theorem (Informal)

∇H Lipschitz continuous, S compact with finite diameter.
Let $\{\boldsymbol{\theta}_n\}$ a sequence generated by the Algorithm, where $\widetilde{\nabla} H$ and the step size satisfy some assumptions.
Then, we have a sublinear convergence rate of the duality gap:

$$g_n^* \leq \max(c_1 n^{-\frac{1}{2}}, c_2 n^{-1})$$

with appropriate constants c_1 and c_2.
$g_n^* = \min_{0 \leq i \leq n-1} -\nabla H(\boldsymbol{\theta}_i)^\top d_i^{FW}$, d_i^{FW} direction obtained by the Frank-Wolfe algorithm with exact gradient.

Synthetic Datasets

(a) Informative (b) Noisy (c) Complementary

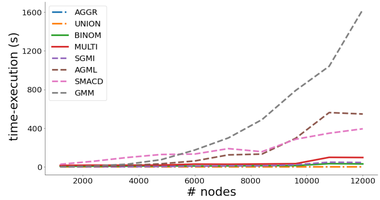
	AGGR	UNION	BINOM	MULTI	SGMI	AGML	SMACD	GMM
APR	0.87	0.89	0.98	0.97	0.78	0.61	0.50	0.86

Performance Ratio
$r_{a,d} = \frac{\mathcal{A}_{a,d}}{\max\{\mathcal{A}_{a,d} \text{ over all } a\}}$

Real World Datasets

		3sources	BBC	BBCSport	Wikipedia	UCI	cora	citeseer	dkpol	aucs	APR
AGGR		0.79	0.91	0.92	0.51	0.95	0.69	0.65	0.73	0.85	
	(+1)	0.74	0.89	0.86	0.42	0.96	0.57	0.53	0.62	0.81	0.90
	(+2)	0.07	0.88	0.80	0.39	0.96	0.49	0.47	0.58	0.77	
UNION		0.75	0.90	0.92	0.51	0.92	0.69	0.65	0.69	0.85	
	(+1)	0.66	0.87	0.85	0.42	0.92	0.57	0.53	0.60	0.65	0.86
	(+2)	0.60	0.84	0.77	0.38	0.92	0.48	0.46	0.55	0.52	
BINOM		0.75	0.88	0.92	0.62	0.97	0.74	0.66	0.62	0.85	
	(+1)	0.76	0.87	0.91	0.57	0.97	0.63	0.59	0.54	0.81	0.94
	(+2)	0.72	0.87	0.90	0.56	0.97	0.64	0.61	0.45	0.77	
MULTI		0.74	0.86	0.88	0.64	0.96	0.76	0.65	0.76	0.85	
	(+1)	0.73	0.87	0.87	0.62	0.96	0.76	0.63	0.72	0.81	0.98
	(+2)	0.75	0.83	0.87	0.59	0.96	0.74	0.63	0.69	0.77	
SGMI		0.75	0.76	0.84	0.61	0.94	0.72	0.51	0.31	0.75	
	(+1)	0.58	0.76	0.83	0.59	0.94	0.72	0.52	0.31	0.76	0.84
	(+2)	0.57	0.76	0.64	0.59	0.94	0.72	0.52	0.31	0.76	
SMACD		0.62	0.69	0.73	0.24	0.33	0.34	0.36	0.26	0.58	
	(+1)	0.60	0.66	0.60	0.25	0.30	0.28	0.32	0.24	0.56	0.56
	(+2)	0.61	0.65	0.78	0.23	0.33	0.37	0.26	0.20	0.60	
GMM		0.80	0.87	0.88	0.57	0.93	0.69	0.58	0.63	0.81	
	(+1)	0.76	0.84	0.77	0.47	0.93	0.58	0.49	0.34	0.77	0.85
	(+2)	0.71	0.81	0.73	0.43	0.93	0.55	0.45	0.27	0.76	

Time-execution (Synthetic Datasets)



References

BINOM-MULTI: Venturini, S., Cristofari, A., Rinaldi, F., and Tudisco, F. (2023). Learning the Right Layers: a Data-Driven Layer-Aggregation Strategy for Semi-Supervised Learning on Multilayer Graphs. Accepted to International Conference on Machine Learning (ICML).
SGMI: Karasuyama, M. and Mamitsuka, H. (2013). Multiple graph label propagation by sparse integration. IEEE transactions on neural networks and learning systems, 24(12):1999–2012.
AGML: Nie, F., Li, J., Li, X. (2016). Parameter-free auto-weighted multiple graph learning: a framework for multiview clustering and semisupervised classification. In IJCAI, pages 1881–1887.
SMACD: Gujral, E. and Papalexakis, E. E. (2018). SMACD: Semi-supervised multi-aspect community detection. In Proceedings of the 2018 SIAM International Conference on Data Mining, pages 702–710. SIAM.
GMM: Mercado, P., Tudisco, F., and Hein, M. (2019). Generalized matrix means for semi-supervised learning with multilayer graphs. Advances in neural information processing systems, 32.