Achieving Optimal Misclassification Proportion in Stochastic Block Model - Review

Author: Chao Gao, Zongming Ma, Anderson Y.Zhang, Harrison H.Zhou

Jiaxin Hu

First edition: 04/19/2020; Final edition: 04/21/2020

ABSTRACT

This paper proposes a polynomial time two-stage method to detect the community in network matrix data with stochastic block model(SBM). One stage is to initialize the partition with the provided new greedy clustering method or any other methods with weak convergence condition. The other stage is to refining the node-wise partition via penalized local maximum likelihood estimation. Statistical guarantees for refinement scheme and initialization are given. The paper reviews related literature closely, which provides a good view of network community detection.

1 PROBLEM FORMULATION & MODEL

1.1 *Model*

Let $A \in \{0,1\}^{n\times n}$ be the symmetric adjacency matrix where $A_{ii} = 0, i \in [n]$ and $A_{uv} = A_{vu}, \forall u > v \in [n]$. Assume there are k communities among n nodes. The stochastic block model can be written as below:

$$A_{uv} \sim_{i.i.d} Ber(P_{uv}), \forall u \neq v; \quad \mathbb{E}(A_{uv}) = P_{uv} = B_{\sigma(u)\sigma(v)},$$

where $B \in [0,1]^{k\times k}$ is the connectivity matrix and $\sigma : [n] \to [k]$ is the label function. In matrix form, the model can also be presented as below:

$$\mathbb{E}(A) = P = H^T B H; \quad A = H^T B H + \mathcal{E},$$

where $H \in \{0,1\}^{k\times n}$ is the membership matrix corresponding to the partition σ and $\mathcal{E}$ is a sub-Gaussian noise matrix. This model can be considered as the order-2 case of TBM(M.Wang 2019) and the single layer case of TSBM(J.Lei 2020).

This main goal for this paper is to find an estimate of partition $\hat{\sigma}$ to achieve the optimal minimax misclassification proportion established in Zhang and Zhou(2015).

1.2 *Discussion of Loss measure*

First, the loss measure in Gao's and Zhang's paper is stated below:

$$l(\hat{\sigma}, \sigma) = \min_{\pi\in S_k} \frac{1}{n} \sum_{u\in[n]} \mathbb{I}\{\hat{\sigma}(u) \neq \pi(\sigma(u))\},$$

where S_k stands for the symmetric group on $[k]$ consisting of all permutations of $[k]$ and $\pi(\sigma(\cdot))$ refers a permuted σ. Recall the misclassification proportion in TBM and TBSM:

$$MCR(\hat{M}_k, M_{k,true}) = \max_{r\in[R_k], a\neq a'\in[R_k]} \min\{D^{(k)}_{a,r}, D^{(k)}_{a',r}\}, \text{ where } D^{(k)}_{r,r'} = \frac{1}{d_k}\sum_{i=1}^{d_k} \mathbb{I}[m^{(k)}_{i,r} = \hat{m}^{(k)}_{i,r'} = 1], r, r' \in [R_k].$$

To show the relationship between $l(\hat{\sigma}, \sigma)$ and $MCR(\hat{M}_k, M_{k,true})$, define:

$$D_{r,r'} = \sum_{u\in[n]} \mathbb{I}\{\hat{H}_{ur} = H_{ur'} = 1\}; \quad D^{\pi}_{r,r'} = \sum_{u\in[n]} \mathbb{I}\{\hat{H}_{ur} = \pi(H_{ur'}) = 1\}$$

For here is an order-2 symmetric case, there is only one confusion matrix D for TBM. Therefore, we can show

$$MCR(\hat{H}, H) = \frac{1}{n} \max_{a\neq a'\in[k], r\in[k]} \min\{D_{ar}, D_{a'r}\}.$$

Let $n_r = \sum_{u\in[n]} \mathbb{I}\{\pi(\sigma(u)) = r\}, r \in [k]$ and $\pi(H)$ refers to the membership matrix corresponds to $\pi(\sigma)$. Then we have,

$$\begin{aligned}
\sum_{u\in[n]} \mathbb{I}\{\hat{\sigma}(u) \neq \pi(\sigma(u))\} &= \sum_{n\in[n]} \sum_{r=1}^{k} \mathbb{I}\{\hat{H}_{ur} = 0, \pi(H_{ur}) = 1\} \\
&= \sum_{r=1}^{k} \sum_{n\in[n]} \mathbb{I}\{\hat{H}_{ur} = 0, \pi(H_{ur}) = 1\} \\
\text{Because } \sum_{u\in[n]} \mathbb{I}\{\pi(H_{ur}) = 1\} = n_r, \ &= \sum_{r=1}^{k} \left(n_r - \sum_{u\in[n]} \mathbb{I}\{\hat{H}_{ur} = 1 = \pi(H_{ur})\} \right) \\
&= \sum_{r=1}^{k} (n_r - D^{\pi}_{r,r}) \\
&\geq^{*} \sum_{r=1}^{k} (n_r - \max_{r'\in[k]} D_{r',r}) \\
&\geq \sum_{r=1}^{k} (\max_{a,a'\in[k]} \min\{D_{ar}, D_{a'r}\}) \\
&\geq \max_{a,a'\in[k], r\in[k]} \min\{D_{ar}, D_{a'r}\},
\end{aligned}$$

where the inequality $\geq^{*}$ is not necessarily become equality even though $\pi = \pi^0 = \arg\min_{\pi\in S_k} \sum_{u\in[n]} \mathbb{I}\{\hat{\sigma}(u) \neq \pi(\sigma(u))\}$. Take a toy example. Let $k = 3$ and suppose below confusion matrix:

$$D = \begin{bmatrix} 10 & 9 & 8 \\ 5 & 6 & 6 \\ 5 & 5 & 6 \end{bmatrix}.$$

Here $D^{\pi_0} = D$ while $D^{\pi_0}_{r,r} < \max_{r'\in[3]} D_{r',r}$ for $r = 2, 3$. We can also calculate $MCR(\hat{H}, H) = \frac{6}{60}$ while $l(\hat{\sigma}, \sigma) = \frac{5+5+9+5+8+6}{60}$. We can conclude that $l(\hat{\sigma}, \sigma) \geq MCR(\hat{H}, H)$. And when the clusters k is fixed, there are $l(\hat{\sigma}, \sigma) \asymp MCR(\hat{H}, H)$.When $k \to \infty$, for $l(\hat{\sigma}, \sigma)$:

$$l(\hat{\sigma}, \sigma) = \frac{\sum_{r\neq r', r, r'\in[k]} D^{\pi}_{r,r'}}{n} = \frac{k(k-1)\bar{D}^{\pi}_{r,r'}}{n},$$

where $\bar{D}^{\pi}_{r,r'}$ is the average of $D^{\pi}_{r,r'}, \forall r \neq r' \in [k]$, which has value at the same level of the entry in D. Here I am wonder how does the entry of D change along with k, n. Consider the worst case of random guess when the community sizes are nearly equal, $D_{r,r'} \asymp \frac{n}{k^2}$. From this perspective, $MCR \asymp \frac{D_{r,r'}}{n} \asymp \frac{1}{k^2}$ will goes to infinity as long as $k \to \infty$ while $l(\hat{\sigma}, \sigma) = O(1)$ when $k \to \infty, n \to \infty$ for the worst case.

Similarly, for the $\eta = MCR = \frac{1}{n_{min}} \max_{a \neq a' \in [k], r \in [k]} \min\{D_{ar}, D_{a'r}\}$ in Lei's paper, consider the nearly equal community sizes case, $n_{min} \asymp \frac{n}{k}$. We still have $\eta \asymp \frac{D_{r,r'}k}{n} \asymp \frac{1}{k}$. That also implies η in Lei's paper will vanish when $k \to \infty$ even though the worst case.

1.3 *Discussion of Optimal Misclassification proportion*

The optimal minimax misclassification proportion is established by Zhang and Zhou(2015).