修士論文 要旨

次数修正確率ブロックモデルにおけるベイズ推定

数理情報学専攻 48166225　新田 猛

指導教員　清 智也 准教授

1　はじめに

本論文では, ネットワークと呼ばれるグラフ型のデータが与えられたときに, ネットワークのノードをいくつかのコミュニティに分割するという問題を扱う.

コミュニティ分割の問題では, ネットワークの生成過程にコミュニティの構造を持つ確率モデルを仮定することが多い. その代表例が確率ブロックモデルである. 確率ブロックモデルは同じコミュニティ内のノードが確率的に同等であり, 例えば次数の期待値が等しいという性質をもつが, この性質のために実社会で現れるネットワークのモデルとしては不適切となる場合がある. この問題に対処するために提案された確率モデルが次数修正確率ブロックモデル [3] である.

本研究では, コミュニティ数が未知の次数修正確率ブロックモデルに対して, MCMC による事後分布からのサンプリングによってコミュニティ割当やコミュニティ数のベイズ推定を行う手法を提案する.

2　次数修正確率ブロックモデル

2.1　定義

本研究におけるネットワークは単純無向グラフとする. ノード数 n, コミュニティ数 K の次数修正確率ブロックモデルは, パラメータ g, B, θ を用いて次のように定義される：

$$A_{ij} \mid g, B, \theta \sim \mathrm{Bernoulli}(B_{g_i g_j}\theta_i\theta_j) \quad (1 \le i < j \le n). \tag{1}$$

ただし, A は無向グラフの隣接行列である. また, $g \in \{1,\dots,K\}^n$ はコミュニティ割当と呼ばれ, 各ノードがどのコミュニティに属するかを表すパラメータである. また, $B \in \mathcal{B} \subset \mathbb{R}_+^{K\times K}$ は正の対称行列, $\theta = (\theta_1,\dots,\theta_n) \in \Theta \subset \mathbb{R}_+^n$ は正のベクトルである. ここで, $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \ge 0\}$ であり, $\mathcal{B}$, Θ は $0 \le B_{g_i g_j}\theta_i\theta_j \le 1$ となることを保証するためのパラメータ空間である.

次数修正確率ブロックモデルは, ラベルの置換およびパラメータ B, θ のある変換に関して確率分布が不変である. モデルの識別可能性のため, ラベルの置換に関する不変性には推定に当たってラベルの置換に対して不変な損失を用いることで対処する. またパラメータ B, θ の変換に関する不変性には, パラメータに制約

$$\sum_{i:g_i=k} \theta_i = n_k \quad (k = 1,\dots,K) \tag{2}$$

を課すことで対処する. ここで, n_k はコミュニティ k に属するノードの個数である.

2.2　既存手法

コミュニティ数が未知の次数修正確率ブロックモデルに対する推定手法には, IDCBM [2] がある. IDCBM では, 無限個のコミュニティが存在することを仮定し, コミュニティ割当の事前分布に中華料理店過程を導入する. しかし, 中華料理店過程とクラスタリングの意味で等価なディリクレ過程は, 混合分布モデルのクラスタ数の推定において一致性を持たないことが示されている [5] ため, 中華料理店過程はコミュニティ数の推定の意味では適切であるとはいえない.

3　提案手法

本研究で提案する手法は, 混合分布モデルで用いられる MCMC の手法である allocation sampler [6] と, allocation sampler を確率ブロックモデルに応用した手法 [4] を基にしている.

allocation sampler を適用するには, モデルのパラメータに関する積分が解析的に計算できなければならないが, 次数修正確率ブロックモデルの定義式 (1) に対してそのような事前分布を導入することは困難である. そこで, グラフに多重辺や自己ループを認めることで次数修正確率ブロックモデルを以下のように近似する：

$$\begin{aligned}
&A_{ij} \mid g, B, \theta \sim \mathrm{Poisson}\left(B_{g_i g_j}\theta_i\theta_j\right) && (1 \le i < j \le n),\\
&A_{ii}/2 \mid g, B, \theta \sim \mathrm{Poisson}\left(B_{g_i g_i}\theta_i^2/2\right) && (i = 1,\dots,n).
\end{aligned}$$

これらの近似に伴い, パラメータ空間 $\mathcal{B}$, Θ をそれぞれ, $\mathcal{B}$ は正の実数を要素に持つ $K \times K$ 対称行列全体, Θ は制約式 (2) を満たす正の n 次元ベクトル全体に拡げる. そして, 近似した次数修正確率ブロックモデルに対して, 事前分布を以下のように設定する：

$$\begin{aligned}
&K \sim \mathrm{Poisson}(1) \mid K > 0,\\
&g_i \mid \pi \sim \mathrm{Multinomial}(1; \pi_1,\dots,\pi_K) && (i = 1,\dots,n),\\
&\pi \mid K \sim \mathrm{Dirichlet}(\alpha),\\
&(\theta)_{g_i=k}/n_k \mid g \sim \mathrm{Dirichlet}(\alpha') && (k = 1,\dots,K),\\
&B_{kl} \mid K \sim \mathrm{Gamma}(k', \theta') && (1 \le k \le l \le K).
\end{aligned}$$

ここで, α, α', k', θ' は正の実数値をとるハイパーパラメータであり, $(\theta)_{g_i=k}$ は θ のうちコミュニティ k に属するノードに対応する成分のみを集めたベクトルを意味する.

このように事前分布を設定すると, 同時分布 $P(A, g, B, \theta, \pi, K)$ のパラメータ B, θ, π に関する積分によって $P(A, g, K)$ が解析的に計算できる. すると,

$$P(g, K \mid A) = P(A, g, K)/P(A) \propto P(A, g, K)$$

であるから, 周辺化した事後分布 $P(g, K \mid A)$ を比例定数を除いて求めることができる. したがって, MCMC の一種であるメトロポリス・ヘイスティングス法を用いることで, 周辺化した事後分布 $P(g, K \mid A)$ からの g と K に関するサンプリングが可能となる.

提案手法では, Nobile and Fearnside [6] が提案した提案分布のうち, 1 つのノードのコミュニティ割当を他のノードのコミュニティ割当を固定した下での条件付き事後確率に従ってサンプリングするギブスサンプリングと, コミュニティの統合または分割を行う AE (absorption/ejection) サンプリングの 2 種類をランダムに組み合わせたメトロポリス・ヘイスティングス法によってサンプリングを行う.

4 数値実験

提案手法と IDCBM [2] を, ノード数 $n = 100$, コミュニティ数 $K = 2$ の人工データに適用した結果が表 1 である. ただし, コミュニティの大きさを 2 通りに設定し, それぞれ 10 個ずつ隣接行列を生成した. ここで, Rand index は $[0, 1]$ の値をとる 2 つのクラスタリングの近さを示す指標であり, 一致するときに最大値 1 となる. 実験では, 真のコミュニティ割当とサンプルとの Rand index の事後平均を計算した.

また, 提案手法と確率ブロックモデルに対する推定手法 [4] を, political blog data [1] と呼ばれるノード数 $n = 1{,}222$, コミュニティ数 $K = 2$ の実データに適用した結果が表 2 である.

いずれの実験においても, 提案手法によって既存手法よりも良い結果が得られていることが確認できる.

5 結論

コミュニティ数が未知の次数修正確率ブロックモデルに対し, 混合分布モデルに用いられる allocation sampler という手法を応用し, 同様の枠組みによる MCMC によってコミュニティ割当やコミュニティ数のベイズ推定を行う手法を提案した. 提案手法を人工データおよび実データに対して適用したところ, 既存の手法よりも良い結果が得られた. ただし, ハイパーパラメータの設定によってはコミュニティの構造を正しく推定できないことがあり, ネットワークに応じて適切なハイパーパラメータを設定する必要がある.

表 1. 人工データへの適用結果. $P(K \mid A)$ は K の事後確率の平均値を, $\hat{K}_{\mathrm{MAP}}$ は K の MAP 推定量によって各 K が推定された回数を, Rand index は事後平均の平均値を表す.

(n_1, n_2)	手法	K	1	2	3	4 以上	Rand index
$(50, 50)$	IDCBM	$P(K \mid A)$	0.000	0.659	0.172	0.169	0.968
		$\hat{K}_{\mathrm{MAP}}$	0	9	1	0	
	提案	$P(K \mid A)$	0.000	0.955	0.043	0.002	0.981
		$\hat{K}_{\mathrm{MAP}}$	0	10	0	0	
$(25, 75)$	IDCBM	$P(K \mid A)$	0.000	0.501	0.315	0.183	0.899
		$\hat{K}_{\mathrm{MAP}}$	0	7	3	0	
	提案	$P(K \mid A)$	0.008	0.933	0.057	0.002	0.960
		$\hat{K}_{\mathrm{MAP}}$	0	10	0	0	

表 2. political blog data [1] への適用結果. $\hat{K}_{\mathrm{MAP}}$ は K の MAP 推定量を, 平均は K の事後平均を表す.

手法	$\hat{K}_{\mathrm{MAP}}$	平均	Rand index
[4]	23	23.36	0.556
提案手法	12	12.36	0.662

参考文献

[1] L. Adamic and N. Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pp. 36–43, New York, NY, USA, 2005. ACM.

[2] T. Herlau, M. Schmidt, and M. Mørup. Infinite-degree-corrected stochastic block model. *Physical Review E*, Vol. 90, 032819, 2014.

[3] B. Karrer and M. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, Vol. 83, 016107, 2011.

[4] A. McDaid, T. Murphy, N. Friel, and N. Hurley. Improved Bayesian inference for the stochastic block model with application to large networks. *Computational Statistics and Data Analysis*, Vol. 60, pp. 12–31, 2013.

[5] J. Miller and M. Harrison. Inconsistency of Pitman–Yor process mixtures for the number of components. *The Journal of Machine Learning Research*, Vol. 15, pp. 3333–3370, 2014.

[6] A. Nobile and A. Fearnside. Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, Vol. 17, pp. 147–162, 2007.