

異種混合メンバーシップ・ブロックモデルと情報推薦への応用

石黒 七海[†] 江口 浩二[†] 横峯 樹[†]

[†] 神戸大学工学部 〒 657-8501 兵庫県神戸市灘区六甲台町 1-1

E-mail: †ishiguro@cs25.scitec.kobe-u.ac.jp, eguchi@port.kobe-u.ac.jp

あらまし ネットワークで表現されたデータの潜在構造をモデル化するアプローチとして、混合メンバーシップ・ブロックモデルが知られている。本研究ではそれを拡張し、複数種類のノードまたはリンクで構成されたネットワークの潜在構造をモデル化する。また、それを用いて、ユーザのアイテム利用履歴に基づいて構成された二部グラフと社会ネットワークなどを合成することによって得られた異種ネットワークに対して適用し、上位 N 件推薦問題に応用し、評価を行う。

キーワード 混合メンバーシップ・ブロックモデル, 異種ネットワーク, 情報推薦

Nanami ISHIGURO[†], Koji EGUCHI[†], and Tatsuki YOKOMINE[†]

[†] Faculty of Engineering, Kobe University

1-1 Rokkodai, Nada, Kobe, Hyogo 657-8501 Japan

E-mail: †ishiguro@cs25.scitec.kobe-u.ac.jp, eguchi@port.kobe-u.ac.jp

1. はじめに

近年、インターネットの普及に伴い、情報の量とそれを扱うユーザの数が爆発的に増加している。これらの多大な量の情報の中から、自分の必要としている情報を選択することは容易ではない。また同時に、ネットワーク構造で表される情報も増加している。その代表的な例が、Facebook や Twitter である。インターネットユーザが増え続けている今、これらのネットワーク構造で表される情報を有効的に使いたいという需要が高まっている。これらの問題を解決するためには、ユーザの嗜好を分析し、ユーザの嗜好にあった情報を推薦するということが必要である。

本稿では、ベイズ的アプローチによるネットワークモデリングに着目する。ベイズ的アプローチによるモデリングには大きく二つの流れがあると考えられる。一つ目は、ブロックモデルと呼ばれるモデルをベイズ的統計学的に拡張した確率論的ブロックモデル (Stochastic Block Models: SB) [4] である。これはノードが単一のクラスタに割り当たるようにクラスタリングし、同じクラスタのノードはすべて相手クラスタにリンクを張る確率が同じであるという仮定のもとにネットワークのモデリングを行ったものである。また、上記で述べたモデルではクラスタの数を所与のパラメータとしている。この問題を回避したモデルとして無限関係モデル (Infinite Relational Model: IRM) [2] がある。IRM は SB において固定的なクラスタの数だけ次元があったベクトルを Chinese Restaurant Process (CRP) によっ

て無限次元に拡張したモデルである。二つ目に、Airoldi らが提案した混合メンバーシップ・ブロックモデル (Mixed Membership Blockmodels: MMSB) [1] がある。このモデルでは、ノードに対する潜在的なグループを確率分布として表現し、グループ間のリンクの生成確率を考慮している。前者は多項分布で表現され、ディリクレ事前分布を仮定する。また、後者はベルヌーイ分布で表現され、これにベータ事前分布を導入することができる。これにより、各ノードが複数のグループに属することを許容する柔軟なモデリングが可能になる。そこで本稿では MMSB のアプローチをとる。

しかし、MMSB ではネットワーク上におけるノードの種類や、リンクに付与されたラベルを直接的に考慮することができない。そこで本稿では、複数の種類のノード、リンクが混在する異種ネットワークを扱うよう MMSB を拡張した異種混合メンバーシップ・ブロックモデル (Heterogeneous Mixed Membership Stochastic BlockModel: HMMB) を提案する。リンクの種類に着目するモデルと、ノードの種類に着目するモデル、リンクに付与された複数のラベルを考慮するモデルを検討し、従来手法に対する優位性を示す。

2. 混合メンバーシップ・ブロックモデル

まずは、本稿で用いる定義についてまとめる。グラフを $G = (N, Y)$ で表し、 N をノード集合、 Y を隣接行列とする。観測されるデータについて、ノード p, q に対応する隣接行列の要素を $Y(p, q) \in \{0, 1\}$ と表す。MMSB において、ある 2 ノー

ド間のリンクはそれぞれのノードの潜在的なグループの多項分布と、それぞれのグループ対における辺の起こりやすさを示すベルヌーイ分布から生成される。すなわち、各ノードはグループ上の多項分布 $Mult(\pi_p)$ で特徴づけられるとし、グループ g に関する多項分布パラメータを $\pi_{p,g}$ とすると、 $\pi_{p,g}$ はノード p がグループ g に属する確率である。このように、それぞれのノードはリンクごとに異なるグループを持つことができる。また、グループ間の関係は $K \times K$ 行列で表されたパラメータで決まるベルヌーイ分布 $Bern(B)$ によって定義される。ここで $B(g, h)$ はグループ g に属するノードから、グループ h に属するノードへの辺が存在する確率であり、 K はグループ数である。指示ベクトル $\mathbf{z}_{p \rightarrow q}$ はノード p からノード q へリンクが存在するとき、ノード p に割り当てられる潜在グループを表している（該当するグループの成分が 1 であり、他が 0 であるベクトル）、それに対し $\mathbf{z}_{p \leftarrow q}$ はノード q に割り当てられる潜在グループを表している。これら二つのベクトルの集合は、それぞれ $\{\mathbf{z}_{p \rightarrow q} \mid p, q \in N\} = Z_{\rightarrow}$ と $\{\mathbf{z}_{p \leftarrow q} \mid p, q \in N\} = Z_{\leftarrow}$ である。以上の定義から、MMSB モデルによるリンクの生成過程を以下に示す。

- (1) すべてのノード p に対して
 - ハイパーパラメータ α で特定されたディリクレ分布から多項分布パラメータ π_p を選択
- (2) すべてのグループの対 (g, h) に対して
 - ハイパーパラメータ $\psi = (\psi_0, \psi_1)$ で特定されたベータ分布からベルヌーイ分布パラメータ $B(g, h)$ を選択
- (3) すべてのノード対 (p, q) に対して
 - 多項分布 $Mult(\pi_p)$ から指示ベクトル $\mathbf{z}_{p \rightarrow q}$ を選択
 - 多項分布 $Mult(\pi_q)$ から指示ベクトル $\mathbf{z}_{p \leftarrow q}$ を選択
 - ベルヌーイ分布 $Bern(\mathbf{z}_{p \rightarrow q}^T B \mathbf{z}_{p \leftarrow q})$ から $Y(p, q)$ を生成
このとき、データ Y と潜在変数 $\pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}$ および B の同時分布（完全同時分布）は以下ようになる。

$$\begin{aligned}
P(Y, \pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}, B | \alpha, \psi) \\
= P(B | \psi) \prod_{p, q, p \neq q} P(Y(p, q) | \mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}, B) P(\mathbf{z}_{p \rightarrow q} | \pi_p) \\
P(\mathbf{z}_{p \leftarrow q} | \pi_q) \prod_p P(\pi_p | \alpha)
\end{aligned} \quad (1)$$

また、完全条件付確率分布は以下ようになる。

$$\begin{aligned}
P(z_{p \rightarrow q} = g, z_{p \leftarrow q} = h | Y, Z_{\rightarrow}, Z_{\leftarrow}; \alpha, \psi) \\
\propto \frac{C(p, g) + \alpha_g}{\sum_g C(p, g) + \sum_g \alpha_g} \frac{C(q, h) + \alpha_h}{\sum_h C(q, h) + \sum_h \alpha_h} \\
\frac{C(g, h, \delta) + \psi_\delta}{C(g, h, 0) + C(g, h, 1) + \psi_0 + \psi_1} \\
= \pi_{p,g} \pi_{q,h} B(g, h)
\end{aligned} \quad (2)$$

ここで、 $C(p, g)$ はあるノード p がグループ g に割り当てられた頻度である。 $C(g, h, 0), C(g, h, 1)$ は、観測されたネットワーク全体におけるノード対において、一方のグループが g であり、他方のグループが h であるときに、リンクが存在する頻度と存在しない頻度をそれぞれ示している。 $\delta \in \{1, 0\}$ とし、リンクの有無を表す。MMSB のグラフィカルモデルを図 1 で示す。

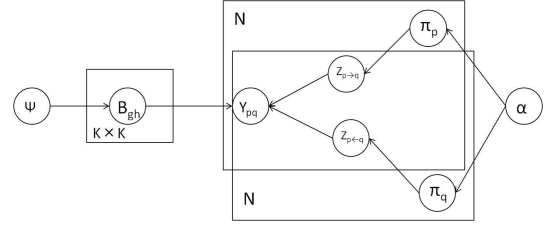


図 1 MMSB のグラフィカルモデル

3. 提案手法

本章では、提案手法である HMMB-L と HMMB-N, LHMMB について述べる。HMMB-L ではリンクの種類に着目し、HMMB-N はノードの種類に着目したモデル、LHMMB はノードの種類とリンクに付与された複数のラベルを考慮するモデルである。

3.1 HMMB-L

MMSB を拡張したモデルである HMMB-L について述べる。MMSB と同様に、用いる定義についてまとめる。グラフを $G = (N, Y, Q)$ と表し、観測されるデータは、ノード p, q について隣接行列の要素を $Y(p, q) = \delta$ とそのリンクの種類を $Q(p, q) = t$ と表す。 $\delta \in \{0, 1\}$ であり、リンクがあるかないかを示すパラメータとする。また、 $t \in \{1, 2, \dots, T\}$ であり、リンクの種類を表すパラメータである。ある 2 ノード間のリンクは、MMSB と同様にして二種類の多項分布から生成される。すなわち、ノード p がグループ g に属する確率 $\pi_{p,g}$ と、グループ g に属するノードから、グループ h に属するノードへの辺が存在する確率 $B(g, h)$ である。さらに、ある種類のリンクのグループ間の関係を、 $K \times K \times T$ の三次のテンソル M で定義されたグループ対毎の多項分布によって表す。ここで K はグループ数であり、 T はネットワーク上に T 種類のリンクの種類が存在することを表す。よって、 $M(g, h, t)$ は t 番目の種類のリンクの種類において、グループ g のノードからグループ h のノードへの辺が存在する確率である。また、指示ベクトル $\mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}$ については、ノード p, q に割り当てられる潜在グループを表している（該当するグループの成分が 1 であり、他が 0 であるベクトル）。これら二つのベクトルの集合は、それぞれ $\{\mathbf{z}_{p \rightarrow q} \mid p, q \in N\} = Z_{\rightarrow}$ と $\{\mathbf{z}_{p \leftarrow q} \mid p, q \in N\} = Z_{\leftarrow}$ である。以上の定義から、HMMB-L モデルによるリンクの生成過程を以下に示す。

- (1) すべてのノード p に対して
 - ハイパーパラメータ α で特定されたディリクレ分布から多項分布パラメータ π_p を選択
- (2) すべてのグループの対 (g, h) に対して
 - ハイパーパラメータ $\psi = (\psi_0, \psi_1)$ で特定されたベータ分布からベルヌーイ分布パラメータ $B(g, h)$ を選択
 - リンクの種類 $t \in \{1, \dots, T\}$ に対して
 - ハイパーパラメータ ϕ で特定されたディリクレ分布から多項分布パラメータ $M(g, h, t)$ を選択
- (3) すべてのノード対 (p, q) に対して

- 多項分布 $Mult(\pi_p)$ から指示ベクトル $\mathbf{z}_{p \rightarrow q}$ を選択
- 多項分布 $Mult(\pi_q)$ から指示ベクトル $\mathbf{z}_{p \leftarrow q}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{p \rightarrow q}^\top B \mathbf{z}_{p \leftarrow q})$ から $Y(p, q)$ を生成
 - $Y(p, q) = 1$ のとき, 多項分布 $Mult(\mathbf{z}_{p \rightarrow q}^\top M \mathbf{z}_{p \leftarrow q})$ から $Q(p, q)$ を生成

このとき, データ Y, Q と潜在変数 $\pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}$ および M, B の同時分布 (完全同時分布) は以下ようになる.

$$\begin{aligned}
 & P(Y, Q, Z_{\rightarrow}, Z_{\leftarrow}, \pi_p, \pi_q, B, M | \alpha, \psi, \phi) \\
 &= \prod_p P(\mathbf{z}_{p \rightarrow q} | \pi_p) P(\pi_p | \alpha) \prod_q P(\mathbf{z}_{p \leftarrow q} | \pi_q) P(\pi_q | \alpha) \\
 & \prod_g \prod_h P(Y(p, q) | \mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}, B) P(B | \psi) \\
 & P(Q(p, q) | \mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}, M) P(M | \phi) \quad (3)
 \end{aligned}$$

また, 完全条件付確率分布は以下ようになる.

$$\begin{aligned}
 & P(z_{p \rightarrow q} = g, z_{p \leftarrow q} = h | Y, Q, Z_{\rightarrow}, Z_{\leftarrow}; \alpha, \psi, \phi) \\
 & \propto \frac{C(p, g) + \alpha_g}{\sum_g C(p, g) + \sum_g \alpha_g} \frac{C(q, h) + \alpha_h}{\sum_h C(q, h) + \sum_h \alpha_h} \\
 & \frac{C(g, h, \delta) + \psi_\delta}{C(g, h, 1) + C(g, h, 0) + \sum_\delta \psi_\delta} \frac{C(g, h, t) + \phi_t}{C(g, h, 1) + C(g, h, 0) + \sum_t \phi_t} \quad (4)
 \end{aligned}$$

ここで, $C(p, g)$ はあるノード p がグループ g に割り当てられた頻度である. $C(g, h, \delta)$ は, 観測されたネットワーク全体におけるノード対において, 一方のグループが g であり, 他方のグループが h である回数であり, δ はリンクがあるかないかを表す. 同様に $C(g, h, t)$ はあるノードがグループ g であり, 相手のノードがグループ h である回数で, t はリンクの種類を表す. HMMB-L のグラフィカルモデルを図 2 で示す.

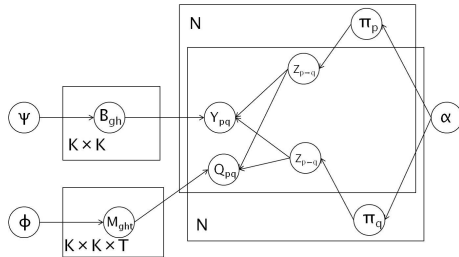


図 2 HMMB-L のグラフィカルモデル

3.2 HMMB-N

続いて, HMMB-N について述べる. 用いる定義についてまとめる. グラフを $G = (N, Y, L)$ と表し, 観測されるデータは, ノード p, q について隣接行列の要素を $Y(p, q) = \delta$ とそのノードの種類を $L(p) = x$ と表す. $\delta \in \{0, 1\}$ であり, リンクがあるかないかを示す. ここでは簡単のため, $x \in \{U, I\}$ であり, それぞれ, ユーザ・ノード, アイテム・ノードを示すとする. ある 2 ノード間のリンクはそれぞれのノードの潜在的なグループの多項分布と, それぞれのグループ対における辺の起こりやすさを表すベルヌーイ分布から生成される. すなわち, 各ノードはグループ上の多項分布 $Mult(\pi_p^x)$ で特徴づけられるとし, x は隣接ノードの種類である. グループ g に関する多項分布パラメータを $\pi_{p,g}^x$ とすると, これはノードの種類が

x であるとき, 任意の種類 x のノード p がグループ g に属する確率である. またグループ間の関係は $K \times K$ 行列で表されたパラメータで決まるベルヌーイ分布 $Bern(B^{x_1, x_2})$ によって定義される. ここで $B^{x_1, x_2}(g, h)$ はグループ g に属する種類 x_1 のノードから, グループ h に属する種類 x_2 のノードへの辺が存在する確率であり, K はグループ数である. また, 指示ベクトル $\mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}$ については, ノード p, q に割り当てられる潜在グループを表している (該当するグループの成分が 1 であり, 他が 0 であるベクトル). これら二つのベクトルの集合は, それぞれ $\{\mathbf{z}_{p \rightarrow q} | p, q \in N\} = Z_{\rightarrow}$ と $\{\mathbf{z}_{p \leftarrow q} | p, q \in N\} = Z_{\leftarrow}$ である. 以上の定義から, HMMB-N モデルによるリンクの生成過程を以下に示す.

(1) 隣接するノードの種類が U であるすべてのノード p に対して

- ハイパーパラメータ α で特定されたディリクレ分布から多項分布パラメータ π_p^U を選択

(2) 隣接するノードの種類が I であるすべてのノード q に対して

- ハイパーパラメータ α で特定されたディリクレ分布から多項分布パラメータ π_q^I を選択

(3) ノードの種類 (U, I) のすべてのグループの対 (g, h) に対して

- ハイパーパラメータ $\psi^{UI} = (\psi_0^{UI}, \psi_1^{UI})$ で特定されたベータ分布からベルヌーイ分布パラメータ $B^{UI}(g, h)$ を選択

(4) ノードの種類 (I, U) のすべてのグループの対 (g, h) に対して

- ハイパーパラメータ $\psi^{IU} = (\psi_0^{IU}, \psi_1^{IU})$ で特定されたベータ分布からベルヌーイ分布パラメータ $B^{IU}(g, h)$ を選択

(5) ノードの種類 (U, U) のすべてのグループの対 (g, h) に対して

- ハイパーパラメータ $\psi^{UU} = (\psi_0^{UU}, \psi_1^{UU})$ で特定されたベータ分布からベルヌーイ分布パラメータ $B^{UU}(g, h)$ を選択

(6) ノードの種類 (I, I) のすべてのグループの対 (g, h) に対して

- ハイパーパラメータ $\psi^{II} = (\psi_0^{II}, \psi_1^{II})$ で特定されたベータ分布からベルヌーイ分布パラメータ $B^{II}(g, h)$ を選択

(7) ノードの種類 (U, I) のすべてのノード対 (u, i) に対して

- 多項分布 $Mult(\pi_u^I)$ から指示ベクトル $\mathbf{z}_{u \rightarrow i}$ を選択
- 多項分布 $Mult(\pi_i^U)$ から指示ベクトル $\mathbf{z}_{u \leftarrow i}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{u \rightarrow i}^\top B_{gh}^{UI} \mathbf{z}_{u \leftarrow i})$ から $Y(u, i) \in \{0, 1\}$ を生成

(8) ノードの種類 (I, U) のすべてのノード対 (i, u) に対して

- 多項分布 $Mult(\pi_i^U)$ から指示ベクトル $\mathbf{z}_{i \rightarrow u}$ を選択
- 多項分布 $Mult(\pi_u^I)$ から指示ベクトル $\mathbf{z}_{i \leftarrow u}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{i \rightarrow u}^\top B_{gh}^{IU} \mathbf{z}_{i \leftarrow u})$ から $Y(i, u) \in \{0, 1\}$ を生成

(9) ノードの種類 (U, U) のすべてのノード対 (u, u') に対して

- 多項分布 $Mult(\pi_u^U)$ から指示ベクトル $\mathbf{z}_{u \rightarrow u'}$ を選択
- 多項分布 $Mult(\pi_{u'}^U)$ から指示ベクトル $\mathbf{z}_{u' \leftarrow u}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{u \rightarrow u'}^\top B_{gh}^{UU} \mathbf{z}_{u' \leftarrow u})$ から $Y(u, u') \in \{0, 1\}$ を生成

(10) ノードの種類 (I, I') のすべてのノード対 (i, i') に対して

- 多項分布 $Mult(\pi_i^I)$ から指示ベクトル $\mathbf{z}_{i \rightarrow i'}$ を選択
- 多項分布 $Mult(\pi_{i'}^I)$ から指示ベクトル $\mathbf{z}_{i' \leftarrow i}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{i \rightarrow i'}^\top B_{gh}^{II} \mathbf{z}_{i' \leftarrow i})$ から $Y(i, i') \in \{0, 1\}$ を生成

このとき、データ Y と潜在変数 $\pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}$ および B の同時分布 (完全同時分布) は以下ようになる。

$$P(Y, L, Z_{\rightarrow}, Z_{\leftarrow}, \pi_p, \pi_q, B | \alpha, \psi) \\ = \prod_{p, x_1} P(\mathbf{z}_{p \rightarrow q} | \pi_p^{x_1}) P(\pi_p^{x_1} | \alpha) \prod_{q, x_2} P(\mathbf{z}_{p \leftarrow q} | \pi_q^{x_2}) P(\pi_q^{x_2} | \alpha) \\ \prod_{g, h, x_1, x_2} P(Y(p, q) | \mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}, B^{x_1, x_2}) P(B^{x_1, x_2} | \psi^{x_1, x_2})$$

また、種類 U のノードと種類 I のノード間のリンクに関する完全条件付確率分布は以下ようになる (他の種類のリンクに関しても同様である)。

$$P(z_{p \rightarrow q} = g, z_{p \leftarrow q} = h | Y, L, Z_{\rightarrow}, Z_{\leftarrow}; \alpha, \psi) \\ \propto \frac{C^I(p, g) + \alpha_g}{\sum_g C^I(p, g) + \sum_g \alpha_g} \frac{C^U(q, h) + \alpha_h}{\sum_h C^U(q, h) + \sum_h \alpha_h} \\ \frac{C^{UI}(g, h, \delta) + \psi_\delta^{UI}}{C^{UI}(g, h, 1) + C^{UI}(g, h, 0) + \sum_\delta \psi_\delta^{UI}} \quad (5)$$

ここで、 $C^I(p, g)$ はノード p の隣接するノードの種類が I の時にグループ g に割り当てられた頻度である。 $C^{UI}(g, h, \delta)$ は、観測されたネットワーク全体において、ノードの種類が U である方のグループが g であり、他方の種類が I のノードのグループが h である回数であり、 δ はリンクがあるかないかを表す。HMNB-N のグラフィカルモデルを図 3 で示す。

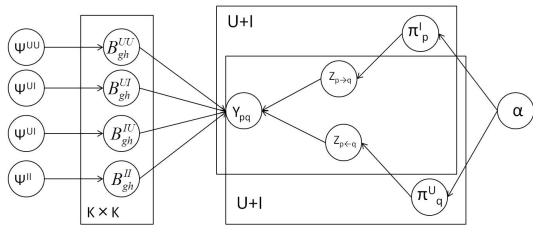


図 3 HMNB-N のグラフィカルモデル

3.3 LHMMB

次に、LHMMB について述べる。用いる定義についてまとめる。グラフを $G = (N, Y, L, R)$ と表し、観測されるデータは、ノード p, q について隣接行列の要素を $Y(p, q) = \delta$ とそのノードの種類を $L(p) = x$ 、任意のリンクに付けられる 1 つまたは複数のラベルを $R(p, q) = \mathbf{t}$ とする。 $\delta \in \{0, 1\}$ であり、リンクがあるかないかを示す。ここでは簡単のため、 $x \in \{U, I\}$ であり、それぞれ、ユーザ・ノード、アイテム・ノードを示すとする。また、ノード p, q 間のリンクにつけられたラベルの数を $S(p, q)$ で

表す。 $\mathbf{t}$ の各成分は $t_s \in \{1, \dots, T\} (s = 1, \dots, S(p, q))$ であり、ラベルの種類を示す。ある 2 ノード間のリンクは HMNB-N と同様にして、2 種類の多項分布から生成される。すなわち、ノードの種類が x であるとき、任意の種類のノード p がグループ g に属する確率 $\pi_{p,g}^x$ と、グループ g に属する種類 x_1 のノードから、グループ h に属する種類 x_2 のノードへの辺が存在する確率 $B^{x_1, x_2}(g, h)$ である。さらに、ある種類のラベルをつけられたリンクのグループ間の関係を $K \times K \times T$ の三次のテンソル M^{x_1, x_2} で表されたパラメータによって決まる多項分布によって定義する。ここで K はグループ数であり、 T はネットワーク上に T 種類のラベルが存在することを表す。よって、 $M^{x_1, x_2}(g, h, t)$ は t 番目の種類のラベルが付けられたノード対において、グループ g に属する種類 x_1 のノードからグループ h に属する x_2 のノードへの辺が存在する確率である。また、指示ベクトル $\mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}$ については、ノード p, q に割り当てられる潜在グループを表している (該当するグループの成分が 1 であり、他が 0 であるベクトル)。これら二つのベクトルの集合は、それぞれ $\{\mathbf{z}_{p \rightarrow q} | p, q \in N\} = Z_{\rightarrow}$ と $\{\mathbf{z}_{p \leftarrow q} | p, q \in N\} = Z_{\leftarrow}$ である。以上の定義から、LHMMB モデルによるリンクの生成過程を以下に示す。

(1) 隣接するノードの種類が U であるすべてのノード p に対して

- ハイパーパラメータ α で特定されたディリクレ分布から多項分布パラメータ π_p^U を選択

(2) 隣接するノードの種類が I であるすべてのノード q に対して

- ハイパーパラメータ α で特定されたディリクレ分布から多項分布パラメータ π_q^I を選択

(3) ノードの種類 (U, I) のすべてのグループの対 (g, h) に対して

- ハイパーパラメータ $\psi^{UI} = (\psi_0^{UI}, \psi_1^{UI})$ で特定されたベータ分布からベルヌーイ分布パラメータ $B^{UI}(g, h)$ を選択
- ラベルの種類 $t \in \{1, \dots, T\}$ に対して

- ハイパーパラメータ ϕ で特定されたディリクレ分布から多項分布パラメータ $M(g, h, t)$ を選択

(4) ノードの種類 (I, U) のすべてのグループの対 (g, h) に対して

- ハイパーパラメータ $\psi^{IU} = (\psi_0^{IU}, \psi_1^{IU})$ で特定されたベータ分布からベルヌーイ分布パラメータ $B^{IU}(g, h)$ を選択

- ラベルの種類 $t \in \{1, \dots, T\}$ に対して
- ハイパーパラメータ ϕ で特定されたディリクレ分布から多項分布パラメータ $M(g, h, t)$ を選択

(5) ノードの種類 (U, U) のすべてのグループの対 (g, h) に対して

- ハイパーパラメータ $\psi^{UU} = (\psi_0^{UU}, \psi_1^{UU})$ で特定されたベータ分布からベルヌーイ分布パラメータ $B^{UU}(g, h)$ を選択

- ラベルの種類 $t \in \{1, \dots, T\}$ に対して
- ハイパーパラメータ ϕ で特定されたディリクレ分布から多項分布パラメータ $M(g, h, t)$ を選択

(6) ノードの種類 (I, I) のすべてのグループの対 (g, h) に

対して

- ハイパーパラメータ $\psi^{II} = (\psi_0^{II}, \psi_1^{II})$ で特定されたベータ分布からベルヌーイ分布パラメータ $B^{II}(g, h)$ を選択

- ラベルの種類 $t \in \{1, \dots, T\}$ に対して
 - ハイパーパラメータ ϕ で特定されたディリクレ分布から多項分布パラメータ $M(g, h, t)$ を選択

(7) ノードの種類 (U, I) のすべてのノード対 (u, i) に対して

- 多項分布 $Mult(\pi_u^I)$ から指示ベクトル $\mathbf{z}_{u \rightarrow i}$ を選択
- 多項分布 $Mult(\pi_i^U)$ から指示ベクトル $\mathbf{z}_{i \leftarrow u}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{u \rightarrow i}^\top B_{gh}^{UI} \mathbf{z}_{i \leftarrow u})$ から $Y(u, i) \in \{0, 1\}$ を生成

- $Y(u, i) = 1$ のとき, $S(u, i)$ のそれぞれについて多項分布 $Mult(\tilde{\mathbf{z}}_{u \rightarrow i}^\top M \tilde{\mathbf{z}}_{i \leftarrow u})$ から $R(u, i)$ を生成

(8) ノードの種類 (I, U) のすべてのノード対 (i, u) に対して

- 多項分布 $Mult(\pi_i^U)$ から指示ベクトル $\mathbf{z}_{i \rightarrow u}$ を選択
- 多項分布 $Mult(\pi_u^I)$ から指示ベクトル $\mathbf{z}_{u \leftarrow i}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{i \rightarrow u}^\top B_{gh}^{IU} \mathbf{z}_{u \leftarrow i})$ から $Y(i, u) \in \{0, 1\}$ を生成

- $Y(i, u) = 1$ のとき, $S(i, u)$ のそれぞれについて多項分布 $Mult(\tilde{\mathbf{z}}_{i \rightarrow u}^\top M \tilde{\mathbf{z}}_{u \leftarrow i})$ から $R(i, u)$ を生成

(9) ノードの種類 (U, U) のすべてのノード対 (u, u') に対して

- 多項分布 $Mult(\pi_u^U)$ から指示ベクトル $\mathbf{z}_{u \rightarrow u'}$ を選択
- 多項分布 $Mult(\pi_{u'}^U)$ から指示ベクトル $\mathbf{z}_{u' \leftarrow u}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{u \rightarrow u'}^\top B_{gh}^{UU} \mathbf{z}_{u' \leftarrow u})$ から $Y(u, u') \in \{0, 1\}$ を生成

- $Y(u, u') = 1$ のとき, $S(u, u')$ のそれぞれについて多項分布 $Mult(\tilde{\mathbf{z}}_{u \rightarrow u'}^\top M \tilde{\mathbf{z}}_{u' \leftarrow u})$ から $R(u, u')$ を生成

(10) ノードの種類 (I, I) のすべてのノード対 (i, i') に対して

- 多項分布 $Mult(\pi_i^I)$ から指示ベクトル $\mathbf{z}_{i \rightarrow i'}$ を選択
- 多項分布 $Mult(\pi_{i'}^I)$ から指示ベクトル $\mathbf{z}_{i' \leftarrow i}$ を選択
- ベルヌーイ分布 $Bern(\mathbf{z}_{i \rightarrow i'}^\top B_{gh}^{II} \mathbf{z}_{i' \leftarrow i})$ から $Y(i, i') \in \{0, 1\}$ を生成

- $Y(i, i') = 1$ のとき, $S(i, i')$ のそれぞれについて多項分布 $Mult(\tilde{\mathbf{z}}_{i \rightarrow i'}^\top M \tilde{\mathbf{z}}_{i' \leftarrow i})$ から $R(i, i')$ を生成

このとき, データ Y, R と潜在変数 $\pi_{1:N}, Z_{\rightarrow}, Z_{\leftarrow}$ および B の同時分布 (完全同時分布) は以下ようになる。

$$\begin{aligned}
 P(Y, L, R, Z_{\rightarrow}, Z_{\leftarrow}, \pi_p, \pi_q, B | \alpha, \psi, \phi) \\
 = \prod_{p, x_1} P(\mathbf{z}_{p \rightarrow q} | \pi_p^{x_1}) P(\pi_p^{x_1} | \alpha) \prod_{q, x_2} P(\mathbf{z}_{p \leftarrow q} | \pi_q^{x_2}) P(\pi_q^{x_2} | \alpha) \\
 \prod_{g, h, x_1, x_2} P(Y(p, q) | \mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}, B^{x_1, x_2}) P(B^{x_1, x_2} | \psi) \\
 P(R(p, q) | \mathbf{z}_{p \rightarrow q}, \mathbf{z}_{p \leftarrow q}, M^{x_1, x_2}) P(M^{x_1, x_2} | \phi)
 \end{aligned}$$

また, 種類 U のノードと種類 I のノード間のリンクに関する完全条件付確率分布は以下ようになる (他の種類のリンクに関しても同様である)。

$$\begin{aligned}
 P(z_{p \rightarrow q} = g, z_{p \leftarrow q} = h | Y, L, R, Z_{\rightarrow}, Z_{\leftarrow}; \alpha, \psi, \phi) \\
 \propto \frac{C^I(p, g) + \alpha_g}{\sum_g C^I(p, g) + \sum_g \alpha_g} \frac{C^U(q, h) + \alpha_h}{\sum_h C^U(q, h) + \sum_h \alpha_h} \\
 \frac{C^{UI}(g, h, \delta) + \psi_\delta^{UI}}{C^{UI}(g, h, 1) + C^{UI}(g, h, 0) + \sum_\delta \psi_\delta^{UI}} \\
 \prod_{s \in S(p, q)} \frac{C^{UI}(g, h, t_s) + \phi_{t_s}^{UI}}{\sum_t C^{UI}(g, h, t) + \sum_t \phi_t^{UI}}
 \end{aligned} \quad (6)$$

ここで, $C^I(p, g)$ はノード p の隣接のノードの種類が I の時にグループ g に割り当てられた頻度である。 $C^{UI}(g, h, \delta)$ は, 観測されたネットワーク全体において, ノードの種類が U である方のグループが g であり, 他方の種類が I のノードのグループが h である回数であり, δ はリンクがあるかないかを表す。同様に $C^{UI}(g, h, t)$ はある種類 U のノードがグループ g であり, 相手のノードが種類 I のグループ g である回数で, t はラベルの種類を表す。また, $t_s (s = 1, \dots, S(p, q))$ は p, q 間のリンクに付けられたラベルのそれぞれの種類を表している。 LHMMB のグラフィカルモデルを図 4 で示す。ただし, ユーザノードとアイテムノードの間のリンクにのみラベルが与えられると仮定している。

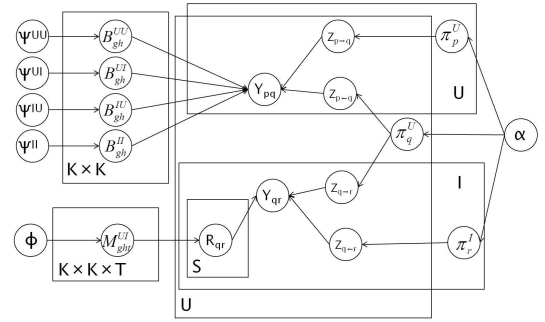


図 4 LHMMB のグラフィカルモデル

4. 実験

提案モデルである HMMB の優位性を示すため, 既存手法との比較実験を行った。

4.1 データセット

本章では, 2つのデータセットを用いて比較実験を行った。

一つ目は, 映画のレビューサイトである Movie Lenz のデータ^(注1)である。MovieLenz のユーザと, ユーザが評価した映画 (アイテム) をノードの種類とする。ユーザ数 943, アイテム数 1682 で, リンク数は 121694 である。ユーザ・アイテム間には, ユーザとそのユーザの評価値の平均値以上に評価しているアイテムにリンクが存在し, アイテム間には, 同じ監督であるか, 同じ俳優が出演している場合にリンクが存在すると仮定し, ユーザ間にはリンクが存在しないと仮定した。このデータにはラベルが存在しないので, HMMB-L, HMMB-N について比較実験を行った。HMMB-L では, ノード対毎にノードの種類の対すなわち $\{UU, UI, IU, II\}$ をリンクの種類とみなした。

(注1) : <http://www.grouplens.org/node/73>

二つ目は、ユーザーの音楽的指向を反映する SNS である Last.fm のデータ^(注2)である。Last.fm のユーザと、ユーザが視聴したアーティスト (アイテム) をノードの種類とする。ユーザ数 1101, アイテム数 524, リンク数 76088 である。ユーザ・アイテム間にはユーザとそのユーザが視聴したアイテムにリンクが存在し、ユーザ間には Last.fm 上で繋がりがああるものにユーザ間リンク存在すると仮定する。ユーザ間には音楽的嗜好が似た者同士にリンクが生まれやすいと考えられる。アイテム間にはリンクが存在しないと仮定した。また、Last.fm 上でユーザはアイテムに複数のラベルを付けることが可能である。このラベルは存在するユーザ・アイテム間リンク上にあると仮定し、ラベルの種類は 250, ラベル数は 32148 個である。HMMB-L, HMMB-N, LHMMB について比較実験を行った。HMMB-L では、ノード対毎にノードの種類の対すなわち $\{UU, UI, IU, II\}$ をリンクの種類とみなした。

4.2 評価方法

ユーザごとにユーザ・アイテム間のリンクを尤度の高いものから順にランキングする上位 N 件推薦問題を想定する。前節のデータを 5 分割し、5-fold cross validation によって評価を行う。

HMMB-L の場合、リンクとその種類に関する尤度は、以下の式で求められる。

$$P(Y | Z_{\rightarrow}, Z_{\leftarrow}, \pi_p, \pi_q, B, M, Q; \alpha, \psi, \phi) = \prod_{p,q} \sum_{g,h} (1 - \rho) \left(\frac{C(p,g) + \alpha_g}{\sum_g C(p,g) + \sum_g \alpha_g} \frac{C(q,h) + \alpha_h}{\sum_h C(q,h) + \sum_h \alpha_h} \frac{C(g,h,\delta) + \psi_\delta}{C(g,h,1) + C(g,h,0) + \sum_\delta \psi_\delta} \frac{C(g,h,t) + \phi_t}{\sum_t C(g,h,t) + \sum_t \phi_t} \right) \quad (7)$$

ここで、 ρ はスパース性のモデリングのためのパラメータ [1] であり、ここでは以下のように定義する。

$$\rho = 1 - d \quad (8)$$

$$d = \sum_{p,q} Y(p,q) / (N \times N) \quad (9)$$

HMMB-N の場合、種類 U のノード p と種類 I のノード q 間のリンクに関する尤度は、以下の式で求められる (他の種類のリンクに関しても同様)。

$$P(Y | Z_{\rightarrow}, Z_{\leftarrow}, \pi_p, \pi_q, B, L; \alpha, \psi) = \prod_{p,q} \sum_{g,h} (1 - \rho^{UI}) \left(\frac{C^I(p,g) + \alpha_g}{\sum_g C^I(p,g) + \sum_g \alpha_g} \frac{C^U(q,h) + \alpha_h}{\sum_h C^U(q,h) + \sum_h \alpha_h} \frac{C^{UI}(g,h,\delta) + \psi_\delta^{UI}}{C^{UI}(g,h,1) + C^{UI}(g,h,0) + \sum_\delta \psi_\delta^{UI}} \right) \quad (10)$$

ここで、 ρ^{UI} は以下のように定義する。

$$\rho^{UI} = 1 - d \quad (11)$$

$$d = \sum_{p,q} Y(p,q) / (U^* \times I^*) \quad (12)$$

U^*, I^* はそれぞれの種類のノード数である。

LHMMB の場合、種類 U のノード p と種類 I のノード q 間のリンクに関する尤度は、以下の式で求められる (他の種類のリンクに関しても同様)。

$$P(Y | Z_{\rightarrow}, Z_{\leftarrow}, \pi_p, \pi_q, B, M, R; \alpha, \psi, \phi) = \prod_{p,q} \sum_{g,h} (1 - \rho^{UI}) \left(\frac{C^I(p,g) + \alpha_g}{\sum_g C^I(p,g) + \sum_g \alpha_g} \frac{C^U(q,h) + \alpha_h}{\sum_h C^U(q,h) + \sum_h \alpha_h} \frac{C^{UI}(g,h,\delta) + \psi_\delta^{UI}}{C^{UI}(g,h,1) + C^{UI}(g,h,0) + \sum_\delta \psi_\delta^{UI}} \prod_{s \in R(p,q)} \frac{C^{UI}(g,h,t_s) + \phi_{t_s}^{UI}}{\sum_t C^{UI}(g,h,t) + \sum_t \phi_t^{UI}} \right) \quad (13)$$

ρ^{UI} は式 (11), (12) に従う。

4.3 パラメータ

HMMB-L については、ハイパーパラメータは α は対称ディリクレを仮定し、その成分を $\alpha = 0.1$ とし、 ψ については、非対称ベータを仮定し、Minka の不動点反復法 [3] に従い、次式で更新する。

$$\psi_\delta^{new} = \psi_\delta \frac{\sum_{g,h} \Psi((1-\rho)C(g,h,\delta) + \psi_\delta) - \Psi(\psi_\delta)}{\sum_{g,h} \Psi(C(g,h,1) + (1-\rho)C(g,h,0) + \sum_\delta \psi_\delta) - \Psi(\sum_\delta \psi_\delta)} \quad \text{if } \delta = 0 \quad (14)$$

$$\psi_\delta^{new} = \psi_\delta \frac{\sum_{g,h} \Psi(C(g,h,\delta) + \psi_\delta) - \Psi(\psi_\delta)}{\sum_{g,h} \Psi(C(g,h,1) + (1-\rho)C(g,h,0) + \sum_\delta \psi_\delta) - \Psi(\sum_\delta \psi_\delta)} \quad \text{if } \delta = 1 \quad (15)$$

HMMB-N, LHMMB については、ハイパーパラメータ α は対称ディリクレを仮定し、その成分を $\alpha = 0.1$ とし、 ψ^{UI} については、非対称ベータを仮定し、Minka の不動点反復法に従い、次式で更新する。

$$\psi_\delta^{UI,new} = \psi_\delta^{UI} \frac{\sum_{g,h} \Psi((1-\rho^{UI})C^{UI}(g,h,\delta) + \psi_\delta^{UI}) - \Psi(\psi_\delta^{UI})}{\sum_{g,h} \Psi(C^{UI}(g,h,1) + (1-\rho^{UI})C^{UI}(g,h,0) + \sum_\delta \psi_\delta^{UI}) - \Psi(\sum_\delta \psi_\delta^{UI})} \quad \text{if } \delta = 0 \quad (16)$$

$$\psi_\delta^{UI,new} = \psi_\delta^{UI} \frac{\sum_{g,h} \Psi(C^{UI}(g,h,\delta) + \psi_\delta^{UI}) - \Psi(\psi_\delta^{UI})}{\sum_{g,h} \Psi(C^{UI}(g,h,1) + (1-\rho^{UI})C^{UI}(g,h,0) + \sum_\delta \psi_\delta^{UI}) - \Psi(\sum_\delta \psi_\delta^{UI})} \quad \text{if } \delta = 1 \quad (17)$$

他の種類の ψ についても同様である。また、グループ数 $K = 10$ として、実験を行った。LHMMB について、ハイパーパラメータ ϕ は対称ディリクレを仮定し、その成分を $\phi = 0.1$ とした。

4.4 結果

本実験では、アイテム推薦のシミュレーションを行うため、訓練データによって推定したモデルによって、ユーザごとにテストデータに含まれる全アイテムをランキングし、以下の 4 つの方法を用いて予測性能を測定した。

- 上位 10 件に含まれる正解アイテムの割合の平均をとる "top-10 precision"

(注2) : <http://www.grouplens.org/node/462>

- 最上位に出現する正解アイテムの順位の逆数の平均をとる”mean reciprocal rank”
- 上位 10 件に少なくとも 1 つの正解アイテムが見つかったユーザの割合である”success rate at 10”
- 正解アイテムが見つかった順位における precision を計算し平均をとる”mean average precision”

MovieLens のデータを用いた結果を図 5, 図 6, 図 7, 図 8 に示す. この実験において, MovieLens のデータにユーザ間リンクがないことから, ”ユーザ間にグループ割り当てをしない”という設定でも実験を行った. これを実験結果上では”HMMB-N(noUU)”と表す. これに対し, HMMB-N はユーザ間リンクが無くてもユーザ間にグループ割り当てを行う.

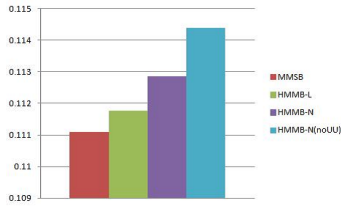


図 5 top-10 precision for MovieLens

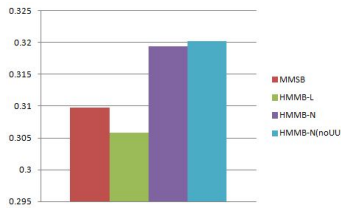


図 6 mean reciprocal rank for MovieLens

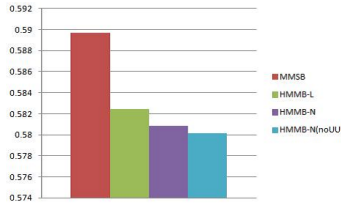


図 7 successrate at 10 for MovieLens

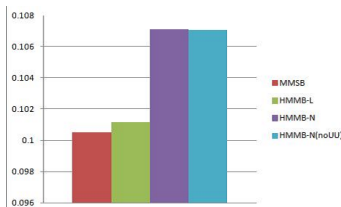


図 8 mean average precision for MovieLens

図 5, 図 6, 図 8 においては, 今回の提案手法である HMMB-N は MMSB や HMMB-L よりも値が上回っており, モデルの精度が改善されているといえる. また, これらの MMSB に対する改善が認められた結果に関して有意性を示すため, Wilcoxon

signed rank test [5] を用いて, MMSB を基準とした有意差の検定を行った. その結果を表 1 に示す. 本検定において, 提案手法である HMMB-N と MMSB との比較をしている.

推薦方法	比較手法	p 値
top-10 precision	HMMB-N	0.1708
	HMMB-N(noUU)	0.1014
mean reciprocal rank	HMMB-N	2.376×10^{-5}
	HMMB-N(noUU)	6.511×10^{-14}
mean average precision	HMMB-N	2.200×10^{-16}
	HMMB-N(noUU)	2.200×10^{-16}

表 1 から, mean reciprocal rank, mean average precision においては, HMMB-N, HMMB-N(noUU) は $p < 0.01$ であるため, MMSB に対して有意に差があることを示した.

Last.fm のデータを用いた結果を図 9, 図 10, 図 11, 図 12 に示す. この実験において, Last.fm のデータにアイテム間リンクがないことから, ”アイテム間にグループ割り当てをしない”という設定でも実験を行った. これを実験結果上では”HMMB-N(noII)”と表す. これに対し, HMMB-N はアイテム間リンクが無くてもアイテム間にグループ割り当てを行う.

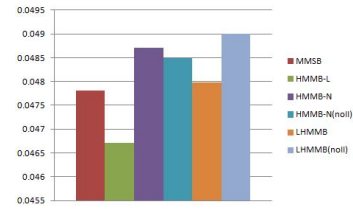


図 9 top-10 precision for Last.fm

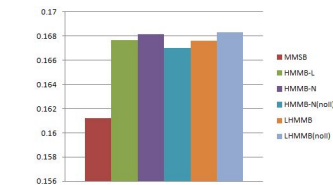


図 10 mean reciprocal rank for Last.fm

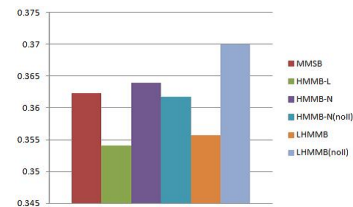


図 11 successrate at 10 for Last.fm

また, 上位 10 件までに出てくる正解アイテムの順位の逆数の平均をとる mean reciprocal rank at 10 も行った. その結果を図 13 に示す.

これらの実験全てにおいて, アイテム間にグループを割り当

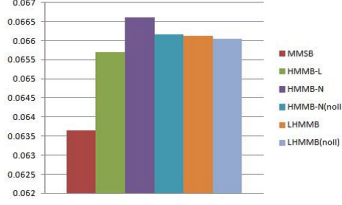


図 12 mean average precision for Last.fm

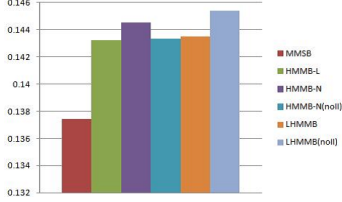


図 13 mean reciprocal rank at 10 for Last.fm

てない LHMMB が MMSB よりも値を上回っている。また、図 10, 図 12, 図 13 においては、今回の提案手法すべてが MMSB を上回っている。また、これらの MMSB に対する改善が認められた結果に関して有意性を示すため、Wilcoxon signed rank test [5] を用いて MMSB を基準とした有意差の検定を行った。その結果の一部の例を表 2 に示す。

表 2 Wilcoxon signed rank test for Last.fm

推薦方法	比較手法	p 値
top-10 precision	HMMB-N	0.08375
	LHMMB(noII)	0.02157
mean reciprocal rank	HMMB-N	0.1789
	LHMMB(noII)	0.005447
mean average precision	HMMB-N	2.200×10^{-16}
	LHMMB	3.163×10^{-11}
mean reciprocal rank at 10	HMMB-N	0.0003202
	LHMMB(noII)	4.956×10^{-6}

表 2 では、mean average precision, mean reciprocal rank at 10 においては、比較したどちらの手法でも $p < 0.05$ であるため、MMSB に対する有意性を確認できた。また、LHMMB, LHMMB(noII) ではすべての評価方法で $p < 0.05$ であり、MMSB に対する有意性を確認できた。

4.5 考 察

MovieLens, Last.fm のどちらデータを用いても、今回の提案手法である HMMB-N が既存手法である MMSB よりも精度の高い推定ができてることがわかる。これは、同一ネットワーク上に異種ノードが存在するとき、ノードの種類を考慮することが有効であるということを示している。表 1, 表 2 において、 $p < 0.05$ である為、MMSB との差は有意であることも示した。ここで、top-10 precision において有意性が見られないのは、top-10 precision にゼロ値が多いという性質上、有意性が現れづらいことが原因であると考えられる。

また、LHMMB(noII) では、一部において HMMB-N よりも高い精度を示した。この現象が top-10 precision, successrate

at 10, mean reciprocal rank at 10 に見られることから、LHMMB は高順位に正解アイテムを推薦するという状況に強いのではないかと考えられる。

一方で、もうひとつの提案手法である HMMB-L の性能は MMSB に及ばないことが多かった。特に、MovieLens のデータを用いた実験では、5-fold cross validation を行った際、一部の fold の結果が不安定になってしまい、平均を引き下げていることが原因である。この不安定性については、HMMB-N や LHMMB に対して、HMMB-L の隣接行列のスパースさが原因であることが考えられる。

様々な評価方法を用いて提案手法を評価したが、提案手法の精度の良さは評価方法によって変化している。これは、各手法がそれぞれの特徴を持っている為であると考えられる。しかし、一部を除き、どの提案手法も概ね MMSB よりも良い精度を示した。

5. おわりに

本論文では、異種ネットワークのための混合メンバシップブロックモデルとして、リンクの種類に着目したモデル HMMB-L と、ノードの種類に着目したモデル HMMB-N、ノードの種類とリンクに対して付与された複数のラベルを考慮したモデル LHMMB を提案し、MovieLens と Last.fm の 2 つのデータを用いて精度の確認を行った。上位 N 件推薦タスクによる評価の結果、HMMB-N は従来の MMSB よりも有効であることを確認した。LHMMB は従来の MMSB よりも概ね有効であったが、さらに一部において HMMB-N よりも高い精度を確認した。一方で、HMMB-L は従来の MMSB の性能に及ばないという状況が多く見られた。より詳細な評価が今後の課題である。

文 献

- [1] E.M. Airoldi, D.M. Blei, S.E. Fienberg, and E.P. Xing. Mixed membership stochastic blockmodels. *The Journal of Machine Learning Research*, Vol. 9, pp. 1981–2014, 2008.
- [2] C. Kemp, J.B. Tenenbaum, T.L. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of the national conference on artificial intelligence*, Vol. 21, p. 381. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006.
- [3] T. Minka. Estimating a dirichlet distribution, 2000.
- [4] K. Nowicki and T.A.B. Snijders. Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, Vol. 96, No. 455, pp. 1077–1087, 2001.
- [5] F. Wilcoxon, S.K. Katti, R.A. Wilcox, Florida State University. Dept. of Statistics, and American Cyanamid Company. *Critical values and probability levels for the Wilcoxon rank sum test and the Wilcoxon signed rank test*. Lederle Laboratories Division, American Cyanamid, 1963.