混合メンバーシップモデルを用いた協調フィルタリングの一検討

横峯 樹[†]　　江口 浩二[‡,††]

[†] 神戸大学工学部情報知能工学科　[‡] 神戸大学大学院工学研究科 情報知能学専攻

1 はじめに

近年，インターネットなどの情報やそれを扱うユーザがますます増加しており，その多くはネットワーク構造として表すことができるため，ネットーワーク分析の需要がますます高くなってきている．また，情報が急激に増加したため代表的な検索技術であるキーワード検索のみではユーザーの嗜好にあった情報を得ることが困難になってきている．そこでユーザごとの履歴情報から嗜好を推定し，そのユーザの嗜好にあった情報を自動的に推薦する仕組みとして情報フィルタリングが注目されている [1]．本稿では，ネットワークに対するノードクラスタリングの手法として知られる Mixed Menbership Stochastic Blockmodels[2](以下:MMSB) を協調フィルタリングに適応しその有効性を示す．さらに，MMSB モデルが二部グラフに限定にしたモデルでないことから，アイテム間にも明示的にリンクが存在するデータを対象とし，その有効性を示す．

2 MMSB モデル

本節では MMSB モデルについて述べる．これはノードに潜在的なグループを割り当て，あるノード対に対してリンクが生成される尤度を推定するモデルである．このモデルは従来の協調フィルタリングのように二部グラフに限定したものではなく，アイテム間のリンクを考慮することが容易である．

定義　まずは本稿の以下で用いる定義についてまとめる．グラフを $G=(N,Y)$ と表し，観測されるデータは隣接行列 $Y(p,q)\in\{0,1\}$ である．ある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分布と，グループ対の分布から生成される．それぞれのノードは $\vec{\pi}_p$ によって特徴づけられ，$\pi_{p,g}$ はノード p がグループ g に属する確率である．つまり，それぞれのノードは複数のグループに属することができる．グループ間の関係はベルヌーイ分布 $B_{K\times K}$ の行列によって定義される．ここで $B(g,h)$ はグループ 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り，K はグループ数を示す．指示ベクトル $\vec{z}_{p\to q}$ 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 (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)，$\vec{z}_{p\leftarrow q}$ はノード q に割り当てられる潜在グループで，これら二つのベクトルの集合はそれぞれ $\{\vec{z}_{p\to q}:p,q\in N\}=Z_{\to}$ と $\{\vec{z}_{p\leftarrow q}:p,q\in N\}=Z_{\leftarrow}$ である．さらに，グループの対，つまり $Y(p,q)$ に対して $(\vec{z}_{p\to q},\vec{z}_{p\leftarrow q})$ は同じである必要はない．これは，非対称なネットワークにも適応可能であることを示している．

MMSB モデル　本稿では各ノードの分布 $\vec{\pi}_p$ をディリクレ事前分布による多項分布で表し，グループ間の分布 $B(g,h)$ をベータ分布を事前分布にとるベルヌーイ分布で表す．

以上の定義より，MMSB モデルによってノードは以下の手順にしたがって生成されると仮定する．

(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 $\vec{\pi}_p$ をサンプリング

(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 $\vec{\phi}=(\phi_1,\phi_2)\in\Phi$ で特定されたベータ分布からベルヌーイ分布 $B(g,h)$ をサンプリング

(3) すべてのノード対 (p,q) に対して
- 多項分布 $\vec{\pi}_p$ から指示ベクトル $\vec{z}_{p\to q}$ をサンプリング
- 多項分布 $\vec{\pi}_q$ から指示ベクトル $\vec{z}_{p\leftarrow q}$ をサンプリング
- $\vec{z}_{p\to q}^{\top}B\vec{z}_{p\leftarrow q}$ から $Y(p,q)$ を生成

この手順のもと，データ Y と潜在変数 $\vec{\pi}_{1:N},Z_{\to},Z_{\leftarrow}$ の結合確率は以下のようになる[1]．

$$\begin{aligned}&P(Y,\vec{\pi}_{1:N},Z^{\to},Z^{\leftarrow},B|\vec{\alpha},\mathbf{\Phi})\\&=P(B|\vec{\phi})\prod_{p,q,p\neq q}P(Y(p,q)|\vec{z}_{p\to q},\vec{z}_{p\leftarrow q},B)\\&\quad P(\vec{z}_{p\to q}|\vec{\pi}_p)P(\vec{z}_{p\leftarrow q}|\vec{\pi}_q)\prod_{p}P(\vec{\pi}_p|\vec{\alpha})\end{aligned}\quad(1)$$

本稿では MMSB モデルの推定方法として，マルコフ連鎖モンテカルロ法の一種であるギブスサンプリングを用いた．

3 MMSB モデルを用いた協調フィルタリング

従来の MMSB モデル　MMSB が最初に提案されたときその目的はネットワーク解析，主にノードのクラスタリングであった [2]．モデルの推定は変分 EM 法で行われており，実験は米国のニューイングランド地方の修道士，ある学校の交友関係ネットワーク，タンパク質の相互作用を対象にクラスタリングを行ったものであった．

MMSB モデルと協調フィルタリング　従来，ノードのクラスタリングなどに使われてきた MMSB モデルであるが，本稿では協調フィルタリングへの適用を試みる．従来の協調フィルタリングでは，アイテムとユーザからなる二部グラフを対象に解析を行うのに対して，ノードのクラスタリングに使われてきた MMSB モデルを用いることで二部グラフ以外のデータに対しての協調フィルタリングを試みる．つまり，ユーザはアイテムだけでなく，他のユーザにもリンクを持つことが可能であり，アイテムはユーザだけでなくアイテム同士でリンクを形成することが可能である．ユーザ同士のリンクは友人や知人と捉えると，社会ネットワークであるとみなせる．アイテム間どうしのリンクは，映画を例にすると，同じ俳優あるいは，同じ監督の映画

A Study on collaborative filtering using of mixed membership models

Tatsuki YOKOMINE[†] and Koji EGUCHI[‡,††], [†]Faculty of Engineering, Kobe University, [‡]Graduate School of Engineering, Kobe University

[1] 文献 [2] では，ベルヌーイ分布のパラメータを直接推定しているが，本稿では推定にギブスサンプリングを適用するため B の事前分布にベータ分布を仮定した

かもしれないし，あるいは同じジャンル，または類似した内容の映画かもしれない．

本稿では従来の協調フィルタリングにはなかったユーザ同士，アイテム同士のリンクを明示的に与えられる MMSB モデルを適用することで，協調フィルタリングの性能を高めることを狙いとする．MMSB モデルを協調フィルタリングに適用した例は，著者らの調査では未だない．

4 実験

本稿ではモデルの正確性を測るための予備検討として，テストセット対数尤度による評価を行う．

データセット データセットとして，映画のレビューサイトである MovieLens[2] のデータを利用した．データはユーザ数 943，アイテム数 1682 の，10 万件の格付けデータであり，各ユーザは少なくとも 20 のアイテムを格付けしている．本稿ではこのデータから二つのデータを作成し，それぞれで MMSB モデルを推定し実験を行った．一つはユーザ，アイテム間にしかリンクが存在しないデータで MovieLens のユーザとアイテムのデータを適用した．二つ目はアイテム間にもリンクが存在するデータで，アイテム間のリンクを考慮する際には MovieLens の分野情報を採用した．具体的に分野は unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western という 19 のフィールドからなっており，それぞれ二値で複数回答を許して評価されている．今回はこの分野が完全に一致し，なおかつ二つ以上の分野に属しているものにリンクを仮定した．この場合，アイテム間のリンク数は 9379 である．

対数尤度による評価 本稿ではアイテム間に明示的なリンクがある場合と，ない場合の二つの場面において MMSB モデルをテストセット対数尤度で評価した．テストセット対数尤度はテストセットパープレキシティの負の対数に比例する．テストセットパープレキシティは言語モデルなどの統計モデルの精度を測定するためによく知られた尺度である [3]．

また，元データの 90 %をトレーニングデータとし，10 %をテストデータとし，90 %でモデルを推定，10 %で評価を行った．なお，アイテム間のリンクはトレーニングデータのみに追加した．

パラメータの設定と実行環境 本稿では，ディリクレ分布のハイパーパラメータ $\alpha = 1.0$，ベータ分布のハイパーパラメータ $\phi_1 = 1.0$，$\phi_2 = 1.0$ とした．ベータ分布のハイパーパラメータはどのグループ対にリンクが生成されやすいかの重み付けの役割を持っている．協調フィルタリングはユーザとアイテムからなるネットワークの解析であり，ユーザとアイテムは違うグループに属すことが多いとみなせるため，偏りを持たせることが考えられるが，本稿ではアイテム間にリンクを含ませることで協調フィルタリングの性能の向上を目的としているため，一様分布を仮定した $\phi_1 = 1.0$，$\phi_2 = 1.0$ で実験を行った．またグループ数 K は 10，30 とし，ギブスサンプリングの繰り返し回数はテストセット対数尤度の改善率が確実に 0.1 %以下になる 500 回とした．

評価結果 表 1 にグループ数 $K = 10, 30$ それぞれにおいてアイテム間にリンクがある場合とない場合でのテストセット対数尤度の値を示し，図 1 にギブスサンプリングの繰り返し回数とテストセット対数尤度の値の遷移を示す．

表 1 および図 1 におけるテストセット対数尤度は辺ごとの期待値である．

表 1: それぞれのトピック数での対数尤度の値

K	アイテム間リンクなし	アイテム間リンクあり
10	-3.972207	-3.966433
30	-3.988724	-3.988311

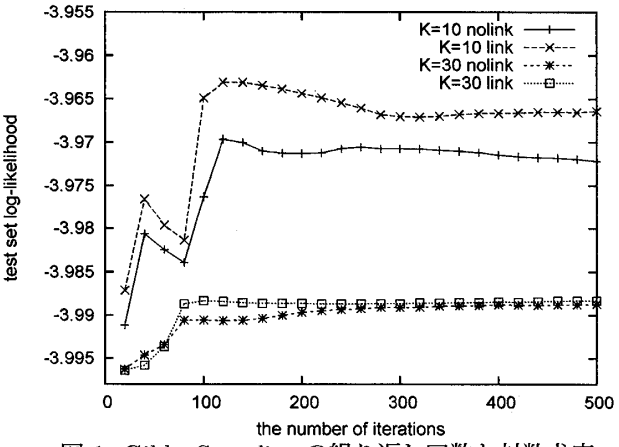

図 1: Gibbs Sampling の繰り返し回数と対数尤度

表 1，図 1 により，K=30 の場合よりも K=10 の場合の方がテストセット対数尤度が良く，K=10 のときにアイテム間リンクを考慮することでテストセット対数尤度が 8 %程度改善していることが確認される．現在，より詳細な実験および分析を行っている．

本稿では協調フィルタリングにおいて新しい手法を提案した．提案手法のタスクに基づく評価，最適なハイパーパラメータの推定，他手法との比較なども今後の課題である．

謝辞 本研究の一部は，科学研究費補助金基盤研究 (B) (20300038) の援助による．

参考文献

[1]Kamishima, T.: 推薦システムのアルゴリズム，人工知能学会誌 22 巻 6 号， Vol. 12, pp. 826–837 (2007).

[2]Edoardo M.Airoldi, David M.Blei, S. E. and P.Xing, E.: Mixed membership stochastic block models, *The Journal of Machine Learning Research*, Vol. 9, pp. 1981–2014 (2008).

[3]Lawrence Rabiner, B.-h. J.: 音声認識の基礎，NTT アドバンステクノロジ株式会社 (1995).

[2] http://grouplens.org/