アイテム評価値の高低を考慮した
混合メンバーシップ・ブロックモデルによる推薦システム

1X07C011-8　井沢祐介
指導教員　後藤正幸

1　研究背景・目的

近年 EC サイトにおけるアイテム数の増加やユーザ嗜好の多様化に伴い，各ユーザの好みに合ったアイテムを自動で推薦するシステムの重要性が高まっている．これは，ユーザの購買履歴やアイテムの評価がデータとして蓄積される EC サイトのマーケティングツールとして，すでに多くのサイトで実装されつつある．

一方，ユーザがアイテムを購買したり評価したりする関係はネットワークとしてモデリングすることが可能であり，このモデルを推薦システムに応用することが可能である．横峯ら [1] は，既にネットワーク分析モデルの一つである混合メンバーシップ・ブロックモデル (以下 MMSB)[2] を推薦システムに適用している．しかし MMSB の性質上，ユーザとアイテムの関係はリンクを持つか持たないかの二値に限定される．推薦システムのデータとして重要なユーザのアイテム評価情報は通常多値であるが，[1] の方法では評価の高低を考慮した推薦を行うことができない．また，ネットワーク分析を目的にしたモデルをそのまま適用しているため，本来異種であるユーザとアイテムを合わせて同じノードで表現しているという問題もある．

そこで本研究では MMSB での推薦システムに多値データを扱う構造を付与し，かつユーザとアイテムを区別したモデルを提案する．さらに，提案モデルにより適したアイテムのランキング法を示す．提案手法を推薦システムのベンチマークデータに適用し，提案手法の有効性を示す．

2　準備

2.1　推薦システム

推薦システムとは，購買・評価履歴からユーザの嗜好を特定し，アイテムを推薦するシステムのことである．本研究では評価履歴を用いた推薦システムを考える．

ユーザ集合を $\mathcal{U} = \{U_i : 1 \leq i \leq n\}$，アイテム集合を $\mathcal{I} = \{I_j : 1 \leq j \leq m\}$ と定義する．ユーザがアイテムを V 段階評価で v 点の評価をした場合は v, 未評価の場合は 0 をとる評価データの行列を $\boldsymbol{R} = (R_{i,j}), 1 \leq i \leq n, 1 \leq j \leq m$ と定義し，未評価アイテムの中からユーザが好むと想定されるものを予測し，推薦する．

2.2　従来研究

MMSB[2] は図 1 のように構成要素であるノードがクラスに分けられ，クラス間でリンクを持つネットワーク構造のモデルである．各ノードはクラス所属確率に従い各クラスに所属する．ノード同士がリンクする確率は，それぞれが所属するクラス間の結合確率と等しいという仮定に基づいている．ただし，図 1 は各ノードが一つのクラスに固定して所属する例を示しているが，MMSB では一般にリンク先の相手ノードによって各ノードの所属クラスが変わることを許容する．

横峯ら [1] は，ユーザとアイテムを区別せずノードで表し，リンクは評価に相当するとして，MMSB の推薦システムへの適用を試みている．ただし，[1] では評価されたデータを評価値に関わらず全て 1，評価されない場合を 0 の二値とし，

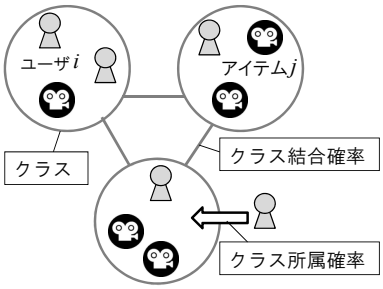

図 1. クラス所属確率とクラス結合確率

ユーザが次に評価しそうなアイテムのみを予測して推薦を行っている．しかし，一般に評価されやすいアイテムと高評価になりやすいアイテムが一致するとは限らないため，高評価アイテムが優先的に推薦されないという問題がある．また，ユーザとアイテムを区別せず同じノードとしてクラスタリングし，ユーザやアイテム同士のリンク，同じクラス内のリンクも許容したモデルとなっている．クラスを分ける基準となるユーザの嗜好とアイテムの特性は異質なものであることに加え，ユーザ同士・アイテム同士の評価データは通常存在しないため，これらの設定は推薦システムへの適用にあたり改善の余地があると考えられる．

3　提案手法

3.1　設定

従来手法の問題点を改善するため，以下では

- 評価値を多値のまま扱う
- ユーザとアイテムに対し，別々のクラスを仮定する
- 同じクラス内のリンクを許容しない

という新たな構造を持つ MMSB モデルを提案する．

3.2　モデルの構造

ユーザとアイテムを別々にクラスタリングするため，ユーザクラス集合を $\mathcal{C}_{\mathcal{X}} = \{C_1^X, C_2^X, ..., C_K^X\}$, アイテムクラス集合を $\mathcal{C}_{\mathcal{Y}} = \{C_1^Y, C_2^Y, ..., C_H^Y\}$，ユーザ U_i とアイテム I_j のリンク時にユーザが所属するクラスを $X_{i,j}$, アイテムが所属するクラスを $Y_{j,i}$ とする．以下，$U_i \in \mathcal{U}, I_j \in \mathcal{I}, C_k^X \in \mathcal{C}_{\mathcal{X}}, C_h^Y \in \mathcal{C}_{\mathcal{Y}}, v \in \{1, 2, ..., V\}$ とし，$X_{i,j}, Y_{j,i}$ を要素とするユーザとアイテムの所属クラス行列を $\boldsymbol{X} = (X_{i,j}), \boldsymbol{Y} = (Y_{j,i})$ と表記する．ユーザ U_i がクラス C_k^X に所属する確率を $\theta_{i,k}^X$，$\theta_{i,k}^X$ を要素とするユーザのクラス所属確率行列を $\boldsymbol{\theta}^X = (\theta_{i,k}^X)$, 同様にアイテムのクラス所属確率行列を $\boldsymbol{\theta}^Y = (\theta_{j,h}^Y)$ とし，クラス C_k^X とクラス C_h^Y の結合確率を $\eta(C_k^X, C_h^Y), \boldsymbol{\eta} = (\eta(C_k^X, C_h^Y))$ とする．データが多値であるため，クラス C_k^X とクラス C_h^Y がリンクするときに v 点が付与される確率として $\phi_v(C_k^X, C_h^Y)$，$\boldsymbol{\phi} = (\phi_v(C_k^X, C_h^Y))$ を追加する．$\boldsymbol{\theta}^X, \boldsymbol{\theta}^Y, \boldsymbol{\phi}, \boldsymbol{\eta}$ の事前分布として，それぞれハイパーパラメータが $\boldsymbol{T}^X = (T_k^X), \boldsymbol{T}^Y = (T_h^Y)$, $\boldsymbol{F} = (F_v(k,h))$ のディリクレ分布と $\boldsymbol{E} = (E_1, E_2)$ のベータ分布を仮定する．リンクには評価値とリンクがないの場合の 0 を含めた $R_{i,j} \in \{0, 1, 2, ..., V\}$ が付与されている．このとき，提案手法のデータとすべての変数，パラメータの同時分布は (1) 式で与えられ，グラフィカルモデルは図 2 のようになる．

$$P(\boldsymbol{R},\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\theta}^X,\boldsymbol{\theta}^Y,\boldsymbol{\eta},\boldsymbol{\phi}|\boldsymbol{T}^X,\boldsymbol{T}^Y,\boldsymbol{E},\boldsymbol{F}) = P(\boldsymbol{R}|\boldsymbol{X},\boldsymbol{Y},\boldsymbol{\phi},\boldsymbol{\eta})P(\boldsymbol{X}|\boldsymbol{\theta}^X)P(\boldsymbol{Y}|\boldsymbol{\theta}^Y)\cdot P(\boldsymbol{\theta}^X|\boldsymbol{T}^X)P(\boldsymbol{\theta}^Y|\boldsymbol{T}^Y)P(\boldsymbol{\eta}|\boldsymbol{E})P(\boldsymbol{\phi}|\boldsymbol{F}) \quad (1)$$

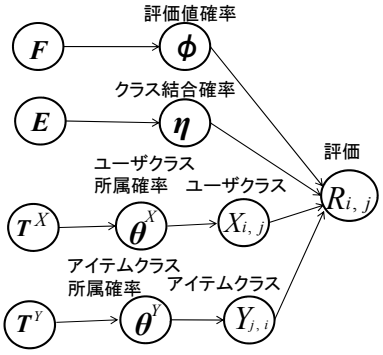

図 2. 提案手法のグラフィカルモデル

3.3 学習・予測アルゴリズム

提案手法の学習・予測アルゴリズムを以下に示す.

Step1) 各ハイパーパラメータを用いて，$\boldsymbol{\theta}^X,\boldsymbol{\theta}^Y,\boldsymbol{\eta},\boldsymbol{\phi}$ の初期値を生成する.

Step2) $\boldsymbol{\theta}^X,\boldsymbol{\theta}^Y,\boldsymbol{\eta},\boldsymbol{\phi}$ の現在値から以下の式でギッブスサンプリングを行って $\boldsymbol{X}$ の事後分布を近似し，この分布に従いユーザのクラスタリングを行う．$\boldsymbol{X}_{i,-j}$ を $\boldsymbol{X}$ の全ての要素のうち $X_{i,j}$ を除いたものと定義すると，$\boldsymbol{X}$ の事後分布は

$$P(X_{i,j}=C_k^X|\boldsymbol{R},\boldsymbol{\theta}^X,\boldsymbol{\eta},\boldsymbol{\phi},\boldsymbol{X}_{i,-j},\boldsymbol{Y}) = Q_{i,j}\theta_{i,k}^X\eta(C_k^X,Y_{j,i})\phi_{R_{i,j}}(C_k^X,Y_{j,i}) \quad (2)$$

で与えられる．$Q_{i,j}$ は i,j には依存するが，k には依存しない基準化定数である．$\boldsymbol{Y}$ の事後分布も式 (2) と同様に求める.

Step3) Step2 の結果から各ディリクレ分布のハイパーパラメータを更新し，再び $\boldsymbol{\theta}^X,\boldsymbol{\theta}^Y,\boldsymbol{\eta},\boldsymbol{\phi}$ を生成する.

Step4) Step2,3 を繰り返し，値が収束したら以下の提示条件 $A_{i,j}$ が大きい順にアイテムを推薦する.

$$A_{i,j}=\sum_k\sum_h\{\theta_{i,k}^X\theta_{j,h}^Y\eta(C_k^X,C_h^Y)\sum_v v\phi_v(C_k^X,C_h^Y)\} \quad (3)$$

□

(3) 式の $\eta(C_k^X,C_h^Y)\sum_v v\phi_v(C_k^X,C_h^Y)$ は評価値の条件付き期待値であり，ユーザに評価自体がされやすく，かつ評価された時に高評価となりやすいアイテムを判別できる．ただし，この条件付き期待値はクラスの組み合わせに依存する．条件付き期待値のみを用いて従来と同様にユーザへ未評価アイテムを推薦する場合では，2.2 節で述べた MMSB の仮定より，ユーザと未評価アイテムを組み合わせごとにそれぞれが持つクラス所属確率の分布に従ってクラスに所属させ，そのクラスの組み合わせの条件付き期待値を用いることになる．このため，ユーザまたは未評価アイテムのクラス所属確率が低いものの条件付き期待値が高いクラスの組み合わせ，つまりユーザの嗜好と合致しない特性の未評価アイテムも推薦されてしまうという問題点がある．このような問題を改善するため，ユーザとアイテムのクラス所属確率 θ_k^X,θ_h^Y を混合比として全てのクラスの組み合わせで混合することで，ユーザとアイテムの組み合わせごとの予測評価値を算出することができる.

4 実験

提案手法の有効性を示すため，推薦システムのベンチマークデータで推薦アイテムの予測実験を行い，提案手法の推薦精度の評価を行う.

4.1 実験条件

実験には，Movielens の映画評価データ 10 万件を使用する．ユーザ数 $n=943$，アイテム数 $m=1682$ である．各ユーザは最低でも 20 個以上のアイテム (映画) を 5 段階で評価している．10 万件のデータをランダムに学習用の 8 万件とテスト用の 2 万件に分け実験を行う，という作業を 5 回繰り返す．(3) 式の結果が大きい順に N 件をユーザに推薦し，推薦したアイテムがテストデータに含まれ，かつ高評価である割合を表す Top-N 精度で評価する.

提案手法の設定に関して，ユーザクラス数 $K=10$，アイテムクラス数 $H=20$ とする．ハイパーパラメータは従来にならい，$\boldsymbol{T}^X,\boldsymbol{T}^Y,\boldsymbol{E},\boldsymbol{F}$ の要素を全て 1.0 に設定する．また，ギッブスサンプリングの繰り返し回数は 2000 回とした.

精度の比較として，従来手法に加え評価値確率 $\boldsymbol{\phi}$ のみを用いて推薦を行う方法を比較手法として用いる.

4.2 実験結果と考察

従来手法・比較手法・提案手法で 5 回ずつ実験を行い，$N=1,5,10$ に対する Top-N 精度の平均をとった結果を図 3 に示す．これより，全ての N に対して提案手法の精度が最も高いことが分かる．また，比較手法の精度が従来手法に勝っていることから，多値の評価データをそのまま扱い，評価値確率 $\boldsymbol{\phi}$ で評価値の高低を考慮したモデルにすることで精度が向上することが示された．ただし提案手法と比較手法の精度から，クラス結合確率 $\boldsymbol{\eta}$ の情報も推薦精度の向上に対して有用であると考えられる.

提案手法の精度が向上した原因として，従来手法と比較手法では推薦の性質が異なることが挙げられる．従来手法では人気のあるアイテムを推薦できるが，実際に利用したユーザには評価値が低いアイテムも推薦してしまう．一方，比較手法では高評価のアイテムを推薦できるが，一部のユーザにしか支持されないニッチなアイテムも推薦してしまう．従って $\boldsymbol{\eta}$ と $\boldsymbol{\phi}$ の両方を取り入れ，アイテムの人気と評価値のバランスを考えた提案手法が最も精度が高くなったと考えられる.

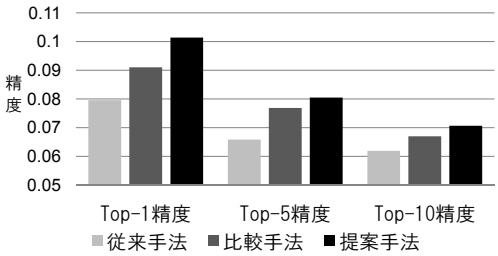

図 3. 各手法による Top-N 精度

5 まとめと今後の課題

本研究では MMSB を推薦システムへ適用する際の問題点に着目し，多段階の評価値を扱うことができユーザとアイテム別のクラスを持つモデルを提案した．さらにそのモデルに適したアイテムのランキング法を提案し，実験によりその有効性を示した.

今後の課題として，ユーザとアイテムの適切なクラス数を推定するノンパラメトリックベイズの導入があげられる.

参考文献

[1] 横峯 樹, 江口 浩二, “ 混合メンバーシップ・ブロックモデルを用いた協調フィルタリング,” 情報処理学会研究報告,No. 6, ROMBUNNO.FI-98,12, 2010.

[2] Edoardo.M.Airoldi, David.M.Blei, Stephen.E.Fienberg, Eric.P.Xing, “Mixed membership stochastic blockmodels,” *Journal of Machine Learning Research,* , 9, pp. 1981-2014, 2008.