評価値とタグ情報の共起を表現する潜在クラスモデルによる協調フィルタリング

1X14C024-1　大堀 祐一
指導教員　後藤 正幸

1　研究背景・目的

近年，EC サイト上では大量の商品が扱われるようになり，ユーザの嗜好も多様化している．これに伴い，自動で各ユーザの嗜好に合致した商品を提示する推薦システムの重要性が高まっている．推薦システムの一つに，各ユーザの評価履歴に基づき，類似した他のユーザとの関係性から評価値予測を行うことでアイテムを推薦する協調フィルタリングがある．ここで，一般にユーザやアイテムの背景には観測できない共通の性質があり，それらを推定しながらクラスタリングする潜在クラスモデルが協調フィルタリング手法として有用であると知られている．

一方，推薦は一般に評価履歴データに基づいて行われるが，実際の EC サイト上では評価値に加えて，タグやレビューをアイテムに付与可能な場合がある．ここで，タグはアイテムに対して強い印象を持った時に付けられるものであり，タグも付与されているアイテムの評価値データは，他のアイテムよりも一貫した意図をもって付けられていると考えられる．そのため，タグが評価値データと共に存在するかどうかを考慮することで評価値予測の精度向上が可能であると考えられる．また，ユーザにとってのタグとアイテムにとってのタグでは意味合いが異なると考えられる．そこで，ユーザ，アイテムのそれぞれに潜在クラスを仮定し，その組み合わせによって評価値予測を行う Flexible Mixture Model[1] に着目し，タグの付与情報をモデルに組み込むことで，評価値予測の精度向上を図る．

また，実際にユーザがアイテムに評価値を付ける際，評価値を付ける前にタグを付けるとは考えにくい．すなわち，評価値予測の際，通常はタグ付与の有無の情報を用いることができない．一方で，タグの付与がある評価値データは一貫した意図をもって付けられており，その評価値は信頼性が高いと考えられる．そこで，本研究では，評価履歴データを活用することで，新規データにタグが付けられるか否かを考慮した予測手法を提案する．具体的には，一度でもタグが付けられたことのあるアイテム（以下，タグ経験アイテム）に対しては，タグが付けられる可能性があり，かつ信頼性が高いデータを予測することが可能であると考えられるため，タグが付けられる前提で予測評価値を算出する．これに対して，一度もタグを付けられたことのないアイテム（以下，タグ未経験アイテム）においては，タグが付けられる可能性が低いとして，タグが付けられない前提で予測評価値を算出する．

以上から，本研究では評価値予測の精度向上のために，評価履歴とタグ付与履歴の双方を用いた潜在クラスモデルを提案する．さらに，対象問題を考慮した評価値予測法も提案する．最後に，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　Flexible Mixture Model[1]

Flexible Mixture Model(以下，FMM) は，ユーザとアイテムのそれぞれに潜在クラスを仮定する確率的潜在クラスモデルである．FMM は，ユーザの評価傾向とアイテムの被評価傾向をそれぞれ別の潜在クラスでクラスタリングし，その組み合わせによって評価値予測を行うモデルである．

ここで，K 個のユーザ潜在クラス集合を $\mathcal{Z} = \{z_k : 1 \leq k \leq K\}$，$L$ 個のアイテム潜在クラス集合を $\mathcal{W} = \{w_l : 1 \leq l \leq L\}$，$I$ 人のユーザ集合を $\mathcal{X} = \{x_i : 1 \leq i \leq I\}$，$J$ 個のアイテム集合を $\mathcal{Y} = \{y_j : 1 \leq j \leq J\}$，評価値 $r \in \{r_1, r_2, ..., r_R\}$ と定義する．このとき，FMM の確率モデルは式 (1) で表される．

$$P(x_i, y_j, r) = \sum_{k=1}^{K}\sum_{l=1}^{L} P(z_k)P(w_l)P(x_i|z_k)P(y_j|w_l)P(r|z_k, w_l) \tag{1}$$

$P(z_k)$，$P(w_l)$，$P(x_i|z_k)$，$P(y_j|w_l)$，$P(r|z_k, w_l)$ は，EM アルゴリズムによって推定する．

また，新規データに対する予測評価値は式 (2) を用いて求める．ここで，$\hat{P}(x_i, y_j, r)$ は EM アルゴリズムにより推定されたパラメータを用いて，計算される同時確率である．

$$\hat{r}_{i,j} = \sum_r r \frac{\hat{P}(x_i, y_j, r)}{\sum_r \hat{P}(x_i, y_j, r)} \tag{2}$$

3　提案手法

3.1　着想

本研究では，評価履歴データに加えて，タグ付与履歴データを適切に扱うことで予測精度の向上を目指す．ここで，ユーザがアイテムに対してタグを自由に付与できる状況を考えると，タグはアイテムに対して強い印象を持った時に付けられるものであるため，タグも付与されている評価値データは他の評価値よりも一貫した意図をもって付けられていると考えられる．そのため，評価値の決定において，タグ付与の有無は重要なファクターとなっていると考えられる．そこで，タグ情報の有無を考慮したモデルを提案し，学習ができれば，予測精度が向上すると考えられる．その際，タグの付与とアイテムの評価傾向に関係性があると考え，アイテムの潜在クラス $\mathcal{W}$ にタグ付与の有無を事象として組み込むこととする．

3.2　タグ情報の有無を考慮した FMM

提案するタグ情報の有無を考慮したモデル（以下，FMMt とする）は FMM に対してアイテムとタグ情報，評価値の組み合わせに潜在クラスを仮定した潜在クラスモデルである．ここで，タグ付与の有無を表す変数を t（ユーザがアイテムに対してタグを付けていれば 1，付けていなければ 0）と定義する．このとき，FMMt の確率モデルは式 (3) で表される．

$$P(x_i, y_j, r, t) = \sum_{k=1}^{K}\sum_{l=1}^{L} P(z_k)P(w_l)P(x_i|z_k)P(t|w_l)P(y_j|w_l)P(r|z_k, w_l) \tag{3}$$

提案モデルのグラフィカルモデルを図 1 に示す．

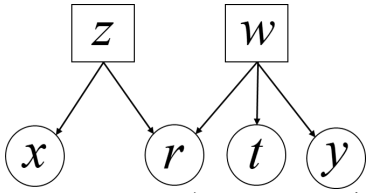

図 1. FMMt のグラフィカルモデル

3.3　パラメータ推定アルゴリズム

提案モデル FMMt のパラメータは潜在変数を含むため，EM アルゴリズムを用いて推定する．ここで，N 件の評価履歴データのうち n 番目のデータで出現するユーザを $a_n \in \mathcal{X}$，アイテムを $b_n \in \mathcal{Y}$，評価値を $c_n \in \{r_1, r_2, ..., r_R\}$，タグ付

与の有無を $d_n \in \{0,1\}$ とする．このとき，対数尤度関数 LL は以下の式 (4) によって計算される．

$$LL = \sum_{n=1}^{N} \log P(a_n, b_n, c_n, d_n) \tag{4}$$

各パラメータは，以下の E-step と M-step を繰り返し，式 (4) により定義された対数尤度関数 LL が収束するまで更新する．

E-step)

$$P(z_k, w_l | a_n, b_n, c_n, d_n) = \frac{P(z_k)P(w_l)P(a_n|z_k)P(b_n|w_l)P(c_n|z_k, w_l)P(d_n|w_l)}{P(a_n, b_n, c_n, d_n)} \tag{5}$$

M-step)

$$P(z_k) = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L} P(z_k, w_l | a_n, b_n, c_n, d_n)}{N} \tag{6}$$

$$P(w_l) = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} P(z_k, w_l | a_n, b_n, c_n, d_n)}{N} \tag{7}$$

$$P(x_i|z_k) = \frac{\sum_{n=1}^{N}\sum_{l=1}^{L} P(z_k, w_l | a_n, b_n, c_n, d_n)\delta(a_n = x_i)}{NP(z_k)} \tag{8}$$

$$P(y_j|w_l) = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} P(z_k, w_l | a_n, b_n, c_n, d_n)\delta(b_n = y_j)}{NP(w_l)} \tag{9}$$

$$P(t|w_l) = \frac{\sum_{n=1}^{N}\sum_{k=1}^{K} P(z_k, w_l | a_n, b_n, c_n, d_n)\delta(d_n = t)}{NP(w_l)} \tag{10}$$

$$P(r|z_k, w_l) = \frac{\sum_{n=1}^{N} P(z_k, w_l | a_n, b_n, c_n, d_n)\delta(c_n = r)}{\sum_r \sum_{n=1}^{N} P(z_k, w_l | a_n, b_n, c_n, d_n)\delta(c_n = r)} \tag{11}$$

ただし，$\delta(\alpha = \beta)$ は $\alpha = \beta$ ならば 1，そうでなければ 0 となるインジケータ関数である．

3.4 評価値予測

ここで，実際にユーザがアイテムに評価値を付ける際，タグの機能は評価値の機能よりもマイナーであるため，評価値を付ける前にタグを付けるとは考えにくい．一方で，タグの付与がある評価値データは一貫した意図をもって付けられており，その評価値は信頼性が高いと考えられる．そこで，式 (12) のように，タグ経験アイテムに対してはタグが付けられる可能性があり，信頼性が高いデータを予測することが有用であると考えられるため，タグが付けられる前提で評価値を予測する．これに対して，タグ未経験アイテムはタグが付けられる可能性が低いとして，タグが付けられない前提で評価値を予測する．

$$\hat{r}_{i,j} = \begin{cases} \sum_r r \frac{\hat{P}(x_i, y_j, r, 1)}{\sum_r \hat{P}(x_i, y_j, r, 1)} & (y_j \text{がタグ経験アイテムの場合}) \\ \sum_r r \frac{\hat{P}(x_i, y_j, r, 0)}{\sum_r \hat{P}(x_i, y_j, r, 0)} & (y_j \text{がタグ未経験アイテムの場合}) \end{cases} \tag{12}$$

4 実験

提案手法の有効性を検証するため，従来手法である FMM と評価値の予測精度を比較する実験を行う．

4.1 実験条件

データセットには，MovieLens[2] の映画評価データを用いた．評価値は 0.5 から 5 までの 10 段階の値である．このデータセットは 4,604,169 件の映画コンテンツの評価データであり，そのうちタグの付与されているデータは 22,400 件となっている．このデータセットから，それぞれユーザ 1 人につきランダムにアイテム 1 件を抽出したテストデータ 24,110 件と残りの学習データ 4,580,059 件に分割した．この学習データを用いてパラメータを推定し，テストデータに対する予測を行い，予測精度により評価を行う．評価指標には平均絶対誤差 (MAE) を用いた．その式はテストデータ数 N_{test}，テストデータにおける評価値 r_{ij} を用いて，以下のようになる．

$$\text{MAE} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} \delta_{ij} \left| \hat{r}_{ij} - r_{ij} \right|}{N_{test}} \tag{13}$$

ただし，δ_{ij} はテストデータ中に r_{ij} が存在する場合は 1，存在しない場合は 0 の値をとるインジケータ関数である．

ここで，一般的に評価値を付ける前に，タグを付けることは考えにくいが，状況によっては評価値よりも先にタグが付けられている可能性もありうる．そこで，予測対象データにタグの付与情報がある場合は，その情報を用いて，式 (14) で予測する．

$$\hat{r}_{i,j,t} = \sum_r r \frac{\hat{P}(x_i, y_j, r, t)}{\sum_r \hat{P}(x_i, y_j, r, t)} \tag{14}$$

また，各データセット・各手法で最適な潜在クラス数は事前実験により探索的に決定した．

4.2 実験結果と考察

表 1 に実験結果を示す．

表 1．各手法における MAE

	モデル	予測式	MAE
(a) 従来手法	FMM	式 (2)	0.702
(b) 提案手法	FMMt	式 (12)	0.676
(c) タグ情報既知	FMMt	式 (14)	0.703

表 1 より，(a) の従来手法と比較して，(b) の提案手法の方が評価値予測の精度が向上していることがわかる．タグを考慮して学習を行うことに効果があり，評価値予測の精度が向上したと考えられる．

また，(c) を見たとき，(a) とほとんど変わらない結果となった．これは，今回のテストデータにおいては，タグの付けられているデータが極端に少なく，タグの有無による影響が小さかったためであると考えられる．このような場合において，提案する評価値予測法が有効であると考えられる．

5 まとめと今後の課題

本研究では評価値予測の精度向上のために，評価履歴とタグ付与履歴の双方を用いた潜在クラスモデルを提案した．さらに，タグ経験アイテムとタグ未経験アイテムで評価値予測の式を変えることで，さらにタグ情報を活用した予測を実現した．

今後の課題として，本研究ではタグの有無のみを考慮したが，さらにタグの種類や個数を考慮したモデルへの拡張などが挙げられる．

参考文献

[1] L. Si and R. Jin, “Flexible mixture model for collaborative filtering,”Proc. 20th International Conference on Machine Learning (ICML ’03), vol. 2, pp. 704-711,Washington, DC, USA, August 2003.

[2] MovieLens. “http://www.movielens.org/”, 2018/1/18 アクセス.