アイテムの多様性を考慮した確率的クラスタリング手法に基づく評価値予測に関する研究

1X12C042-9　國岡 翔
指導教員　後藤 正幸

1　研究背景と目的

近年，EC サイト等では膨大な数の商品が扱われる様になり，ユーザの嗜好も多様化している．そのため，大量の情報の中からユーザの嗜好に合致する商品を自動的に推薦する推薦システムの必要性が高まっている．推薦システムの代表例として，ユーザ間の評価履歴の類似性から未評価アイテムの評価値を予測することでユーザの嗜好に見合うアイテムを提示する協調フィルタリングがある．

本研究では，ユーザのアイテムに対する評価値予測を目的として Nicola らに提唱された Block Mixture Model(以降 BMM)[1] に着目する．BMM は，ユーザとアイテムそれぞれに潜在クラスを仮定することでユーザの嗜好やアイテムの差異を表現し，未評価アイテムの評価値を予測するモデルである．このモデルでは，各アイテムは全潜在クラスへ確率的に所属しており，全アイテムで潜在クラス数が同数となる．一方，潜在クラス数の増加に伴うモデルの表現力の増加と，モデルの複雑化による過学習はトレードオフの関係にあり，潜在クラス数は汎化能力を決める大きな要因となる．また，汎化能力が最適となる潜在クラス数 (以降，最適な潜在クラス数) はデータ数に依存し，一般にデータ数が多いほど潜在クラス数を多く，より複雑な構造を学習させることができる [2]．しかし，評価履歴ではアイテム間で被評価数のばらつきが大きいため，モデル全体で最適な潜在クラス数を一意に選択しても，アイテム単位で見るとそれが最適な潜在クラス数とはなっていない可能性がある．

そこで本研究では，被評価数の差異に着目し，アイテムごとに最適な潜在クラス数を決定し，不必要な潜在クラスへの所属確率に制約を加えることで，モデルの汎化能力を向上させる方法を提案する．また，ベンチマークデータを用いた評価実験により提案手法の有用性を示す．

2　従来手法

2.1　準備

以下に BMM で用いる変数を定義する．

表 1: 変数の定義

変数	説明
u_m	ユーザを表す変数 $m \in \{1, ..., M\}$
i_n	アイテムを表す変数 $n \in \{1, ..., N\}$
z_k	ユーザの潜在クラスを表す変数 $k \in \{1, ..., K\}$
w_ℓ	アイテムの潜在クラスを表す変数 $\ell \in \{1, ..., L\}$
r_m^n	ユーザ u_m のアイテム i_n に対する評価値
$\mathcal{I}_m$	ユーザ u_m に評価されたアイテムの集合
$\mathcal{U}_n$	アイテム i_n を評価したユーザの集合
c_{mk}	ユーザ u_m の潜在クラス z_k に対する所属確率
$d_{n\ell}$	アイテム i_n の潜在クラス w_ℓ に対する所属確率
π_k	ユーザの潜在クラス z_k の混合比
ψ_ℓ	アイテムの潜在クラス w_ℓ の混合比
μ_k^ℓ,$(\sigma_k^\ell)^2$	潜在クラス z_k へ所属するユーザの潜在クラス w_ℓ 所属のアイテムに対する評価値の平均と分散

BMM は，評価値の確率分布に正規分布を仮定した潜在クラスモデルであり，ユーザとアイテムをクラスタリングするステップ，評価値を予測するステップの２段階からなる．

2.2　Block Classification EM による学習

BMM では，ユーザ，アイテム別に潜在クラスを定めており，潜在クラス間の相互依存を考慮しないモデルであるため，ユーザとアイテムそれぞれの変数に対して交互に EM アルゴリズムを適用する Block Classification EM(以下 CEM) を導入してパラメータを推定する．

CEM は 4 ステップからなる繰り返し学習により尤度を最大化するパラメータを推定する方法である．以下に CEM におけるパラメータ推定式を示す．

[E-Step(ユーザ)]

$$c_{mk} = \frac{\left(\prod_{\ell=1}^{L} \phi(u_0^{(u,\ell)}, u_1^{(u,\ell)}, u_2^{(u,\ell)}; \mu_k^\ell, \sigma_k^\ell)\right) \pi_k}{\sum_{k'=1}^{K}\left(\prod_{\ell=1}^{L} \phi(u_0^{(u,\ell)}, u_1^{(u,\ell)}, u_2^{(u,\ell)}; \mu_{k'}^\ell, \sigma_{k'}^\ell)\right) \pi_{k'}} \quad (1)$$

[M-Step(ユーザ)]

$$\pi_k = \frac{\sum_{m=1}^{M} c_{mk}}{M} \quad (2)$$

$$\mu_k^\ell = \frac{\sum_{m=1}^{M} \sum_{i_n \in \mathcal{I}_m} c_{mk} d_{n\ell} r_m^n}{\sum_{m=1}^{M} \sum_{i_n \in \mathcal{I}_m} c_{mk} d_{n\ell}} \quad (3)$$

$$(\sigma_k^\ell)^2 = \frac{\sum_{m=1}^{M} \sum_{i_n \in \mathcal{I}_m} c_{mk} d_{n\ell} (r_m^n - \mu_k^\ell)^2}{\sum_{m=1}^{M} \sum_{i_n \in \mathcal{I}_m} c_{mk} d_{n\ell}} \quad (4)$$

[E-Step(アイテム)]

$$d_{n\ell} = \frac{\left(\prod_{k=1}^{K} \phi(u_0^{(i,k)}, u_1^{(i,k)}, u_2^{(i,k)}; \mu_k^\ell, \sigma_k^\ell)\right) \psi_\ell}{\sum_{\ell'=1}^{L}\left(\prod_{k=1}^{K} \phi(u_0^{(i,k)}, u_1^{(i,k)}, u_2^{(i,k)}; \mu_k^{\ell'}, \sigma_k^{\ell'})\right) \psi_{\ell'}} \quad (5)$$

[M-Step(アイテム)]

$$\psi_\ell = \frac{\sum_{n=1}^{N} d_{n\ell}}{N} \quad (6)$$

$$\mu_k^\ell = \frac{\sum_{n=1}^{N} \sum_{u_m \in \mathcal{U}_n} c_{mk} d_{n\ell} r_m^n}{\sum_{n=1}^{N} \sum_{u_m \in \mathcal{U}_n} c_{mk} d_{n\ell}} \quad (7)$$

$$(\sigma_k^\ell)^2 = \frac{\sum_{n=1}^{N} \sum_{u_m \in \mathcal{U}_n} c_{mk} d_{n\ell} (r_m^n - \mu_k^\ell)^2}{\sum_{n=1}^{N} \sum_{u_m \in \mathcal{U}_n} c_{mk} d_{n\ell}} \quad (8)$$

ただし，

$$\phi(u_0, u_1, u_2; \mu_k^\ell, \sigma_k^\ell) = (\sigma_k^\ell)^{-u_0} exp\left(\frac{2u_1\mu_k^\ell - u_2 - u_0(\mu_k^\ell)^2}{2(\sigma_k^\ell)^2}\right) \quad (9)$$

であり，u_0, u_1, u_2 は正規分布の十分統計量である．CEM の終了条件は，Hold-out サンプルの精度が悪化した時点とし，そのときのパラメータを評価値予測に用いる．

2.3　推定値による評価値予測

CEM により得られたパラメータ $\hat{c}_{mk}$，$\hat{d}_{n\ell}$，$\hat{\mu}_k^\ell$ を用いてユーザ u_m のアイテム i_n に対する予測評価値 $\hat{r}_m^n$ を式 (10) で求める．

$$\hat{r}_m^n = \sum_{k=1}^{K} \sum_{\ell=1}^{L} \hat{c}_{mk} \hat{d}_{n\ell} \hat{\mu}_k^\ell \quad (10)$$

3　提案手法

3.1　概要

評価履歴において，多くのユーザに評価されるアイテムから，数件しか評価されないアイテムまで混在するが，BMM では全てのアイテムで潜在クラス数は同数となる．しかし，最適な潜在クラス数は被評価数に依存するため，アイテムごとに最適な潜在クラス数は異なると考えられる．従って，仮にモデル全体の最適な潜在クラス数を一意に決定しても，ア

イテム単位で見るとそれが最適な潜在クラス数であるとは限らず，被評価数が少ないアイテムでは過学習による汎化誤差を，被評価数が多いアイテムではモデルの表現力不足による汎化能力の低下を招いている可能性がある．

そこで本研究では，予め大きめの潜在クラス数を設定しておき，被評価数が少ないアイテムに対しては最適な潜在クラス数を決定し，不必要な潜在クラスへの所属を制限することで，過学習を抑制する学習方法を提案する．提案手法により，過学習による汎化誤差の影響が深刻でなくなるため，予め設定するアイテムが所属可能な潜在クラス数の上限の増加も可能となる．結果，被評価数の多いアイテムに対するモデルの表現力不足の解消も同時に行えるため，モデル全体の汎化能力の向上も期待できる．

3.2 提案手法の手順

予め設定するアイテムが所属可能な潜在クラスの上限 (以降，上限潜在クラス数) を $\tilde{L}$，アイテム i_n の被評価数を N_n，全アイテムの平均被評価数を $\bar{N}$ と定義する．また，ユーザの潜在クラス数 K はアイテムの潜在クラス数 L と比較してそれほど精度に影響を及ぼさないため，以降ユーザの潜在クラス数 K は固定として考える．提案手法では，アイテムごとに被評価数 N_n に応じた閾値 $f(N_n)$ を設定し，CEM による E-Step(アイテム) において，各アイテムは $1 \leq \ell \leq f(N_n)$ を満たす潜在クラス w_l にのみ所属確率を付与する．ここで $f(N_n)$ は以下の式 (11) で定義する．

$$f(N_n) = \max\left\{\tilde{L}\left(\frac{N_n}{\bar{N}}\right)^{\beta},\ \alpha\right\} \tag{11}$$

β は指数関数のパラメータ，α は全アイテムが所属する潜在クラス数の下限値を表す．このとき提案手法では，CEM による E-Step(アイテム) である式 (5) は，次式のように書き換えられる．

$$d_{n\acute{\ell}} = \frac{\delta_{n\ell}\left(\prod_{k=1}^{K}\phi(u_0^{(i,k)}, u_1^{(i,k)}, u_2^{(i,k)}; \mu_k^{\ell}, \sigma_k^{\ell})^{\delta_{n\ell}}\right)\psi_{\ell}}{\sum_{\ell'=1}^{L}\delta_{n\ell'}\left(\prod_{k=1}^{K}\phi(u_0^{(i,k)}, u_1^{(i,k)}, u_2^{(i,k)}; \mu_k^{\ell'}, \sigma_k^{\ell'})^{\delta_{n\ell'}}\right)\psi_{\ell'}} \tag{12}$$

$$\delta_{n\ell} = \begin{cases} 1 & 1 \leq \ell \leq f(N_n) \\ 0 & otherwise \end{cases} \tag{13}$$

このとき，被評価数が $\bar{N}$ 未満のアイテムは被評価数に応じて指数関数的に所属潜在クラス数が増加し，被評価数が $\bar{N}$ 以上のアイテムは $\tilde{L}$ 個の全潜在クラスに所属する．

4 実験

提案手法の有効性を検証するため，オリジナルの BMM(以降，従来) と，先に述べた提案を施した BMM(以降，提案) との予測精度を比較する実験を行った．

4.1 実験条件

実験には，GroupLens による映画評価データを用いた．データセットはユーザ数 943，アイテム数 1,682，評価履歴 100,000 件の評価値データであり，評価値は 1 から 5 までの 5 段階である．このデータを 20,000 件のテストデータと，80,000 件の学習データに分割し，学習データの 1 割を Hold-out サンプルとした．

また，予備実験の結果より，式 (11) のパラメータは $\alpha = 10$，$\beta = 4$ とし，ユーザの潜在クラス数 K は 10 で固定し，従来のアイテムの潜在クラス数 L 及び提案の上限潜在クラス数 $\tilde{L}$ を 3〜10, 20, 30, ..., 90 と変化させた．さらに，潜在クラスモデルでは，全体的に評価値を高く，または低く付けやすいといったユーザの評価傾向の違いにより，類似した嗜好を持ちながらも別の潜在クラスに所属し，精度が低下しまう可能性がある．そこで，本研究では従来及び提案に対して，ユーザ u_m の全アイテムに対する評価値の平均値を $\bar{r}_m$ とし，評価値 r_m^n を式 (14) の $\tilde{r}_m^n$ へと置き換えて実験を行った．

$$\tilde{r}_m^n = r_m^n - \bar{r}_m \tag{14}$$

4.2 評価指標

評価指標には，テストデータと予測評価値の平均絶対誤差 (MAE) を用いた．テストデータ数を J，テストデータの評価値を v_m^n とし，η_m^n を，v_m^n が存在する場合は 1，存在しない場合は 0 の値を示すインジケータ関数をとすると，MAE は式 (15) で表される．

$$\mathrm{MAE} = \frac{1}{J}\sum_{m=1}^{M}\sum_{n=1}^{N}|\hat{r}_m^n - v_m^n|\eta_m^n \tag{15}$$

4.3 結果と考察

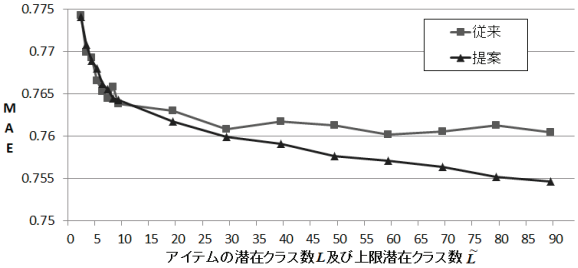

図 1. 潜在クラス数 L, $\tilde{L}$ と MAE の関係

図 1 に潜在クラス数 L, $\tilde{L}$ を変化させたときの予測精度の推移を示す．従来において，潜在クラス数 L が 30 までは，潜在クラスの数増加に伴うモデルの表現力の向上により汎化能力が上昇し続けたが，それ以降では，過学習による汎化誤差も無視できなくなり，両者による影響が均衡し続けたため予測精度が改善しなくなったと考えられる．また，予測精度が劣化せず横ばいが続いていることから，本実験の範囲では，過学習による汎化誤差と併せて，潜在クラス数 L の増加に伴うモデルの表現力の向上も続いていることが分かり，被評価数が多いアイテムの中には，最適な潜在クラス数が 90 以上のアイテムも存在することが示唆される．

一方，提案では被評価数が少ないアイテムに対して最適な潜在クラス数を決定し，それに合わせて所属確率を付与している．その結果，従来では生じていた，被評価数が少ないアイテムの過学習による汎化誤差を抑制しつつ，必要なアイテムに対しては，上限潜在クラス数 $\tilde{L}$ の増加に伴うモデルの表現力の向上が可能となったため，従来手法に比べて汎化能力が高まり, 予測精度も向上し続けたと考えられる．

5 まとめと今後の課題

本研究では，アイテムの被評価数の差異に着目し，アイテムごとに所属できる潜在クラス数を制限した BMM を提案した．また提案手法が従来手法と比べて高精度な評価値の予測ができることを示した．

今後の課題として，潜在クラスへの所属数を決める関数 $f(N_n)$ を，被評価数のみではなく嗜好の分かれ方の大きさにも応じて決定することで，さらに高性能な評価値の予測を実現することが挙げられる．

参考文献

[1] B.Nicola，M.Guarascio，G.Manco，“A Block Coclustering Model for Pattern Discovering in Users Preference Data，”*Communications in Computer and Information Science*，Vol.348, pp 94-108, 2013.

[2] C.M. ビショップ，“パターン認識と機械学習 上，” シュプリンガージャパン, 2008.