所属クラスタ数を考慮したクラスタリング協調フィルタリング手法の提案

1X10C112-6　屋代夢
指導教員　後藤正幸

1　研究背景と目的

情報技術の発達により，EC サイトでは膨大な数のアイテムが扱われている．これらの EC サイトでは，売上を向上させるため，ユーザの嗜好に合致したアイテムを提示する推薦システムを導入している．推薦システムの代表的な手法として，アイテムの評価履歴を用いてユーザの嗜好を予測する協調フィルタリング (以下，CF) がある．CF では，評価履歴が類似しているユーザの情報から被推薦ユーザが好むであろうアイテムの予測を行う．

CF には既に様々な手法が提案されているが，クラスタリングを応用することで，ユーザやアイテムごとに使用する評価履歴データを選別し，評価値の予測精度の向上を図る手法が存在する．このうち，Xu ら [1] はユーザの嗜好やアイテムのジャンルを同時に考慮し，ユーザとアイテムを合わせてクラスタリングし，評価値を予測する手法を考案している．この手法では，クラスタリングを行うためにユーザごと，アイテムごとに各クラスタへの所属確率を定義し，それらを推定するための最適化問題を定式化している．この最適化問題を解くことでクラスタリングを行い，クラスタごとに評価値を予測する．その後，ユーザの所属確率が最大となるクラスタでの予測評価値を基に推薦するアイテムを決定する．この手法では，所属確率が正であれば，そのユーザないしアイテムの評価履歴をクラスタごとの評価値の予測時に使用することになる．ここで，極端に小さな所属確率を認めてしまうと，所属確率が 0 をとることがほとんどなくなり，クラスタごとに評価傾向の差を持たせるクラスタリングができなくなってしまう．これを防ぐため，ユーザやアイテムが所属するクラスタ数を総クラスタ数に対する対数関数で一意に設定している．しかし，この設定により，ユーザやアイテムごとにクラスタ数を適切に変更できず，クラスタリングによる評価履歴の選別が効果的でない可能性が生じている．

そこで本研究は，ユーザごと，アイテムごとに所属できるクラスタ数を適応的に決定し，推薦を行う手法を提案する．具体的には，各ユーザがどのようなアイテム群への興味が強いのかを評価し，様々なアイテム群に興味を持つユーザは所属できるクラスタ数を大きく，少数のアイテム群に興味が集中しているユーザは所属できるクラスタ数が小さくなるように決定を行う．さらに，アイテムについても同様に所属クラスタ数を可変にすることで，ユーザやアイテムごとに適切な所属クラスタ数を与え，予測精度の向上を図る．ベンチマークデータを用いた実験を行い，提案手法の有効性を示す．

2　従来手法

推薦システムで扱うユーザ集合を $\mathcal{U} = \{U_i : 1 \leq i \leq n\}$，アイテム集合を $\mathcal{I} = \{I_j : 1 \leq j \leq m\}$ とする．また，ユーザ U_i がアイテム I_j に対して付けた評価値を T_{ij} とする．ただし，T_{ij} は G 段階評価で g 点の評価をした場合は g，未評価の場合は欠損値をとるものとする．

Xu らの手法は，ユーザのクラスタリング，予測評価値の導出の 2 つのステップから成る．前者は，ユーザとアイテムを同時にクラスタリングするステップであり，後者は，各クラスタで予測評価値を導出し，その中から最終的な予測評価値を決定するステップである．

2.1　ユーザ・アイテムのクラスタリング

いま，クラスタリングにより得られるクラスタ集合を $\mathcal{C} = \{C_k : 1 \leq k \leq h\}$ とし，ユーザ U_i の各クラスタへの所属確率を $\boldsymbol{q}_i = (q_{i1}, \cdots, q_{ih})$，アイテム I_j の各クラスタへの所属確率を $\boldsymbol{r}_j = (r_{j1}, \cdots r_{jh})$ とする．ただし，q_{ik} はユーザ U_i のクラスタ C_k への所属確率，r_{jk} はアイテム I_i のクラスタ C_k への所属確率である．行列 $\mathbf{P} \in \mathbb{R}^{(n+m)\times h}$，$\mathbf{Q} \in \mathbb{R}^{n\times h}$，$\mathbf{R} \in \mathbb{R}^{m\times h}$ をそれぞれ以下で定義し，行列 $\mathbf{P}$ を求める問題へと帰着させる．

$$\mathbf{P} = \begin{pmatrix} \mathbf{Q} & \mathbf{R} \end{pmatrix}^{\mathrm{T}} = \begin{pmatrix} \boldsymbol{q}_1 & \cdots & \boldsymbol{q}_n & \boldsymbol{r}_1 & \cdots & \boldsymbol{r}_m \end{pmatrix}^{\mathrm{T}} \quad (1)$$

Xu らは，行列 $\mathbf{P}$ を求めるための最適化問題 (MCoC) を以下のように定式化した．

$$\underset{\boldsymbol{q}_i, \boldsymbol{r}_j}{\text{minimize}} \ \varepsilon(\mathbf{Q}, \mathbf{R}) = \sum_{i=1}^{n}\sum_{j=1}^{m} \left(\left\| \frac{\boldsymbol{q}_i}{\sqrt{D_{ii}^{row}}} - \frac{\boldsymbol{r}_j}{\sqrt{D_{jj}^{col}}} \right\|^2 T_{ij} \right) \quad (2)$$

subject to

$$\forall i\,, \quad \sum_{k=1}^{h} q_{ik} = 1 \quad (3)$$

$$\forall j\,, \quad \sum_{k=1}^{h} r_{jk} = 1 \quad (4)$$

$$\forall i, j, k\,, \quad q_{ik} \geq 0,\, r_{jk} \geq 0 \quad (5)$$

$$\forall i, j\,, \quad |\boldsymbol{q}_i| = |\boldsymbol{r}_j| = \lceil \log_2 h \rceil \quad (6)$$

ただし，$D_{ii}^{row} = \sum_{j=1}^{m} T_{ij}$，$D_{jj}^{col} = \sum_{i=1}^{n} T_{ij}$ とし，$|\cdot|$ はベクトルの成分のうち 0 でない数，$\lceil A \rceil$ は実数 A 以上の最小の整数を表す．式 (6) により所属クラスタ数を一意に制限している．この最適化問題を解くことにより，ユーザ及びアイテムの各クラスタへの所属確率が求められる．

2.2　予測評価値の導出

最適化問題を解くことで得られる $\boldsymbol{q}_i$，$\boldsymbol{r}_j$ を元に，各クラスタのユーザ集合及びアイテム集合を定義する．各 k に対し，$q_{ik} > 0$ のときユーザ U_i はクラスタ C_k の要素，$r_{jk} > 0$ のときアイテム I_j はクラスタ C_k の要素である．各クラスタのユーザ集合及びアイテム集合を定義することで，各クラスタで使用する評価履歴データが定まるので，ピアソン相関係数によるユーザベース法 [2] により予測評価値の計算を行う．各クラスタで計算される予測評価値のうち，ユーザの所属確率が最も大きいクラスタで計算された値を最終的な予測評価値とする．

3　提案手法

従来手法では，極端に小さな所属確率を割り当てることを防ぎ，クラスタごとに評価傾向が異なるようなクラスタリングが望ましいことから，全てのユーザ及び全てのアイテムが所属可能なクラスタ数を $\lceil \log_2 h \rceil$ (h はクラスタ数) に制約している．これは一定値であるため，ユーザごとの嗜好の違いやアイテムごとのジャンルの違いを考慮した所属クラスタ数の決定がなされていない．したがって，本来は少数のクラスタに所属すべきユーザやアイテム，反対に，本来は多数のクラスタに所属すべきユーザやアイテムが，適切でないクラスタ数を割り当てられる場合があり，使用する評価履歴データの選別が効果的でない可能性がある．

そこで本研究では，ユーザやアイテムごとに所属クラスタ数を変化させることで，評価履歴データをより有効に活用可能とする方法を提案する．式 (6) により，所属クラスタ数は全てのユーザ及び全てのアイテムに対して一定の値に制限さ

れていたが，ユーザやアイテムごとに適切な値を与えることで，MCoC によるクラスタリングの効果を高め，予測精度の向上を図る．

提案手法では，ユーザがどのようなアイテムの集合に興味を持っているかを考慮することで，ユーザごとに異なる所属クラスタ数を割り当てる．類似したアイテムの集合 (以下，アイテム群) を作るため，各ユーザにより付与された評価値を特徴ベクトルとしてアイテムのクラスタリングを行う．ユーザが多くのアイテム群に興味を持っていれば，そのユーザの嗜好が他のユーザの嗜好と部分的に類似している可能性が高く，その評価履歴は多数のユーザに対して有用と考えられるので，所属クラスタ数を大きな値に設定することが望ましい．反対に，興味のあるアイテム群が少なければ，そのユーザの評価履歴は少数のユーザにのみ有用と考えられるので，所属クラスタ数を小さな値に設定することが望ましい．このように所属クラスタ数を決定するため，ユーザのアイテム群への興味の強さを定量化する．定量化した興味の強さを基に，ユーザの所属クラスタ数の決定を行う．以下に，ユーザの所属クラスタ数を決定する手順を示す．アイテムに関しては，ユーザをクラスタリングしてユーザ群を作り，同様の処理を行うことで所属クラスタ数を決定する．

3.1 アイテム群への興味の定量化

まず，アイテム群の形成を行う．各ユーザにより付与された評価値をアイテムの特徴ベクトルとし，$k = h$ とした k 平均法 [3] によるクラスタリングを行いアイテム群を形成する．得られるアイテム群集合を $\mathcal{C}' = \{C'_k : 1 \leq k \leq h\}$ とする．

次に，ユーザ U_i のアイテム群 C'_k に対する興味の強さ σ_{ik} を式 (8) により求める．

$$\sigma_{ik} = \frac{\sum_{j=1}^{m} T_{ij} \cdot \eta_{jk}}{\sum_{j=1}^{m} \eta_{jk}} \tag{7}$$

η_{jk} はアイテム I_j がアイテム群 C'_k に属していれば 1，そうでなければ 0 をとる関数を表す．興味の強さは，ユーザ U_i が持つアイテム群 C'_k に所属するアイテムの評価値の平均を表しており，値が大きいほどそのアイテム群への興味が強いということができる．

3.2 所属クラスタ数の決定

式 (7) で得られた興味の強さを用いて，各ユーザの所属クラスタ数を決定する．式 (8) により，ユーザ U_i の付けた評価値の平均値 $\overline{T}_i$ を計算し，$\overline{T}_i \geq \sigma_{ik}$ となった k の個数を所属クラスタ数として決定する．

$$\overline{T}_i = \frac{\sum_{j=1}^{m} T_{ij}}{\sum_{j=1}^{m} \eta_{ij}} \tag{8}$$

ただし，η_{ij} は，T_{ij} が値を持っていれば 1，そうでなければ 0 をとる関数とする．この決定法により，ユーザが何らかのアイテムを評価していれば所属クラスタ数は 1 から h の間の値をとり，アイテム全体に対して高い評価を与えているほど，所属クラスタ数は大きな値をとる．

以上のようにして所属クラスタ数を決定することで，ユーザの嗜好の偏りを考慮した所属クラスタ数の決定が可能となる．ここで求めた所属クラスタ数を式 (6) の代わりに用い，MCoC を実行し推薦を行う．

4 実験

提案手法の有効性を示すために，推薦システムのベンチマークデータを用いて実験を行い，予測精度の評価を行う．

4.1 実験条件

実験には，MovieLens-100K の映画評価データを用いた．ユーザ数は 943，アイテム数は 1,682 であり，ユーザが視聴した映画の評価が 5 段階評価で与えられている．ユーザの評価履歴数は 10 万件あり，8 万件を学習データ，2 万件をテストデータとしたデータセットを 5 つ作成した．予測精度の評価には MAE を用いるものとした．MAE は次の式 (9) で表される．

$$\mathrm{MAE} = \frac{1}{N} \sum_{t=1}^{N} |y_t - \hat{y}_t| \tag{9}$$

ただし，N は予測評価値とテストデータが共にある数，y_i はテストデータにおける評価値，$\hat{y}_i$ は予測評価値を表す．MAE は予測値と実際の評価値の差異を表すので，値が低いほど精度が高いことを示す．クラスタリングにおけるクラスタ数を 1 から 40 まで変化させて実験を行い，MAE の推移の比較を行った．

4.2 実験結果と考察

図 1 に総クラスタ数 h を変化させて MAE を計算した結果を示す．

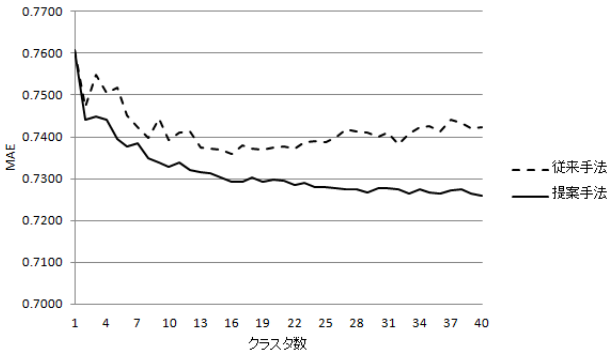

図 1 MAE の比較

図 1 より，提案手法はどのクラスタ数においても従来手法よりも優れていることがわかる．所属クラスタ数をユーザやアイテムごとに適応的に決定することで，従来手法ではノイズであったデータの削減や，活用することが出来なかったデータが利用可能となったために，提案手法の MAE が小さくなったと考えられる．また提案手法では，総クラスタ数が大きな値をとっても，その精度は向上し続けているが，従来手法では精度の劣化がわずかに確認できる．総クラスタ数の増加に伴い，ユーザやアイテムごとに考えられる適切なクラスタ数と，従来手法の $\lceil \log_2 h \rceil$ による所属クラスタ数の差が著しく大きくなり，精度が劣化したと考えられる．一方で，提案手法の場合は，総クラスタ数が増加しても，ユーザやアイテムごとに適切な所属クラスタ数を割り当てることができるので，むしろ精度が向上していると考えられる．

5 まとめと今後の課題

本研究では，クラスタリングを用いた CF において，ユーザやアイテムごとの所属クラスタ数の決定手法の提案を行い，実験によりその有効性を示した．今後の課題として，k 平均法以外のクラスタリング手法による所属クラスタ数の決定の検討などが挙げられる．

参考文献

[1]B. Xu, J. Bin, C. Chen, D. Cai, “An Exploration of Improving Collaborative Recommender Systems via User-Item Subgroups,” *WWW2012 Proc. of the 21st international conference on World Wide Web*, pp.21-30, 2012.

[2] 神嶌敏弘, “推薦システムのアルゴリズム (1), ”人工知能学会誌, Vol.22, No.6, pp.826–837, 2007.

[3]C.M. ビショップ, “パターン認識と機械学習　下, ”スプリンガー・ジャパン, 2007.