目的関数ベースの Rough Membership C-Means クラスタリングに基づく協調フィルタリング

第 7 グループ　KIM HAERANG

1. はじめに

Rough Membership C-Means (RMCM) 法は Hard C-Means (HCM; k-Means) 法にラフ集合理論を導入することで各対象のクラスターに対する帰属の不確実性を取り扱うことができるクラスタリング手法であり，人間の主観的な評価を扱う協調フィルタリング (Collaborative Filtering, CF) において有効であることが報告されている [1]．ただし，既存の RMCM 法はヒューリスティックな手法であるため，本研究では，目的関数を導入することでクラスタリングの妥当性検証や理論的展開を可能とする目的関数ベースの RMCM (RMCM2) 法 [2] に基づく協調フィルタリングを検討した．本研究では，目的関数ベースの RMCM 法に基づく CF (RMCM2-CF) をユーザベースとアイテムベースの 2 つのアプローチから考え，user-based RMCM2-CF および item-based RMCM2-CF を提案した．また，提案法を実データに適用し従来法との比較を行うことで推薦性能を検証した．さらに，目的関数値と推薦性能の関係を観察した．

2. RMCM2 法に基づく協調フィルタリング

2.1. RMCM 法

RMCM 法は，HCM 法をラフ集合理論によって拡張したラフクラスタリングの一種である．RMCM 法では，最近隣割り当てによる暫定クラスターを算出した上で，対象の二項関係 $\mathcal{R}$ による近傍内でのクラスター比率を表すラフメンバシップをクラスターメンバシップとして利用することで帰属の不確実性を取り扱う．

RMCM 法のアルゴリズムを以下に示す．

Step 1 クラスター数 C および二項関係 $\mathcal{R}$ を設定する．

$$R_{it} = \begin{cases} 1 & (||\boldsymbol{x}_t - \boldsymbol{x}_i|| \leq \delta)\,, \\ 0 & (\text{otherwise})\,. \end{cases} \tag{1}$$

ここで，近傍半径 δ は対象間距離分布の τ-パーセンタイルによって決定する．

Step 2 初期クラスター中心 $\boldsymbol{b}_c$ を決定する．

Step 3 対象 i のクラスター c に対するメンバシップ u_{ci} を最近隣割り当てによって求める．

$$u_{ci} = \begin{cases} 1 & \left(c = \underset{1\leq l\leq C}{\arg\min}\, ||\boldsymbol{x}_i - \boldsymbol{b}_l||\right), \\ 0 & (\text{otherwise})\,. \end{cases} \tag{2}$$

Step 4 ラフメンバシップ $\mu_{ci}^{\mathcal{R}}$ を計算する．

$$\mu_{ci}^{\mathcal{R}} = \frac{\sum_{t=1}^{n} R_{it} u_{ct}}{\sum_{t=1}^{n} R_{it}}. \tag{3}$$

Step 5 クラスター中心 $\boldsymbol{b}_c$ を計算する．

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n} \mu_{ci}^{\mathcal{R}} \boldsymbol{x}_i}{\sum_{i=1}^{n} \mu_{ci}^{\mathcal{R}}}. \tag{4}$$

Step 6 u_{ci} に変化がなくなるまで **Step 3-5** を繰り返す．

2.2. RMCM2 法

RMCM 法はヒューリスティックな手法であり，最適化問題として定義されていない．RMCM2 法は，RMCM 法に対し目的関数を導入した手法である．RMCM2 法の目的関数は，RMCM 法と同様のクラスター中心の更新則が導出されるように設計されている．RMCM2 法の目的関数を以下に示す．ただし，$d_{ci} = ||\boldsymbol{x}_i - \boldsymbol{b}_c||$ である．

$$\begin{aligned} J_{\text{RMCM2}} &= \sum_{c=1}^{C}\sum_{i=1}^{n} \mu_{ci}^{\mathcal{R}} d_{ci}^2 \\ \text{s.t.}\quad & u_{ci} \in \{0,1\}, \quad \forall c, i, \\ & \sum_{c=1}^{C} u_{ci} = 1, \quad \forall i. \end{aligned} \tag{5}$$

RMCM2 法のアルゴリズムを以下に示す．

Step 1 クラスター数 C および二項関係 $\mathcal{R}$ を式（1）で設定する．

Step 2 初期クラスター中心 $\boldsymbol{b}_c$ を決定する．

Step 3 対象 i のクラスター c に対するメンバシップ u_{ci} を以下の式より求める．

$$u_{ci} = \begin{cases} 1 & \left(c = \underset{1\leq l\leq C}{\arg\min} \sum_{k=1}^{n} \frac{R_{ki}}{\sum_{t=1}^{n} R_{kt}} d_{lk}^2\right), \\ 0 & (\text{otherwise})\,. \end{cases} \tag{6}$$

Step 4 ラフメンバシップ $\mu_{ci}^{\mathcal{R}}$ を式（3）で計算する．

Step 5 クラスター中心 $\boldsymbol{b}_c$ を式（4）で計算する．

Step 6 u_{ci} に変化がなくなるまで **Step 3-5** を繰り返す．

2.3. user-based RMCM2-CF

user-based RMCM2-CF では，評価値行列 X に RMCM2 法を適用することで嗜好の類似したユーザのクラスターを抽出し，クラスター内で嗜好度の高いアイテムを推薦する．user-based RMCM2-CF の手順を以下に示す．

Step 1 $n \times m$ の評価値行列 $X = \{r_{ij}\}$ に対して RMCM2 法を適用し，$\mu_{ci}^{\mathcal{R}}$ と $\boldsymbol{b}_c$ を求める．

Step 2 ユーザ i に対するアイテム j の推薦度 $\hat{r}_{ij}$ を計算する．

$$\hat{r}_{ij} = \sum_{c=1}^{C} \mu_{ci}^{\mathcal{R}} b_{cj}. \tag{7}$$

Step 3 閾値 $\eta \in [\min\{\hat{r}_{ij}\}, \max\{\hat{r}_{ij}\}]$ 以上の推薦度を持つアイテムを推薦する．

$$\tilde{r}_{ij} = \begin{cases} 1 & (\hat{r}_{ij} \geq \eta)\,, \\ 0 & (\text{otherwise})\,. \end{cases} \tag{8}$$

2.4. item-based RMCM2-CF

item-based RMCM2-CF は，評価値行列の転置行列 $X^\top$ に RMCM2 法を適用することで類似したアイテムのクラスターを抽出し，クラスター内のアイテムを高く評価しているユーザの推薦度を高くする．item-based RMCM2-CF の手順は user-based RMCM2-CF の手順と同様であり，評価値行列の転置行列 $X^\top$ に RMCM2 法を適用する部分だけが異なる．

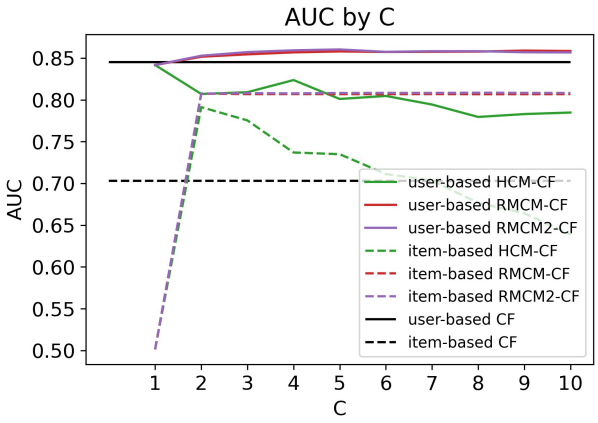

図 1: NEEDS-SCAN/PANEL データにおける初期クラスター数 C による AUC の変化

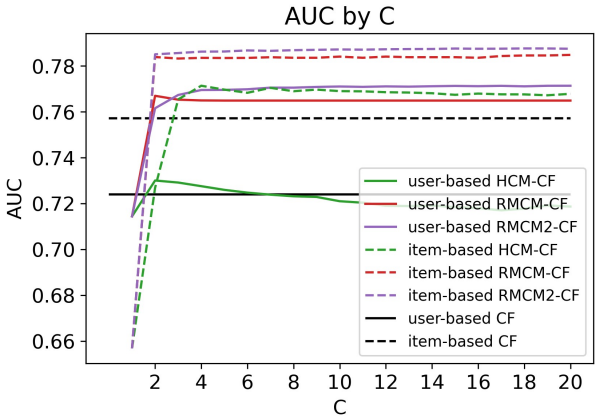

図 2: MovieLens-100k データにおける初期クラスター数 C による AUC の変化

3. 数値実験

3.1. 実験概要

2 つの実データ (NEEDS-SCAN/PANEL, MovieLens-100k) に対して提案法 (user-based RMCM2-CF, item-based RMCM2-CF) を適用し，初期クラスター数 C およびラフさを調節する τ による推薦性能の変化を検証した．比較手法としては，HCM 法に基づく CF (user-based HCM-CF, item-based HCM-CF)，RMCM 法に基づく CF (user-based RMCM-CF, item-based RMCM-CF)，メモリベース CF (user-based CF[3], item-based CF[4]) を用いた．評価指標として ROC-AUC 指標を用いた．また，目的関数値と AUC の値の関係を観察した．

3.2. 実験結果

NEEDS-SCAN/PANEL データにおける初期クラスター数 C による推薦性能の変化を図 1 に示す．RMCM-CF および RMCM2-CF の AUC は τ を種々変化させたときの最大値を採用している．AUC が最大となる τ は user-based RMCM2-CF の場合 $\tau = 10 \sim 40$，item-based RMCM2-CF の場合 $\tau = 50$ となった．図から，user-based と item-based の両方の手法において RMCM-CF および RMCM2-CF が HCM-CF より高い AUC を持つことが確認できる．また，HCM-CF は C が大きくなると性能が低下する傾向を持つ一方，RMCM-CF および RMCM2-CF は C に関わらず高い AUC を維持していることがわかる．全体的な AUC の値を比較すると，user-based 手法の方が item-based 手法より高い値を示した．

MovieLens-100k データにおける初期クラスター数 C による推薦性能の変化を図 2 に示す．AUC が最大となる τ は user-based RMCM2-CF の場合 $\tau = 80 \sim 90$，item-based RMCM2-CF の場合 $\tau = 80$ となった．図から，user-based と item-based の両方の手法において RMCM-CF および RMCM2-CF が HCM-CF より高い AUC を示し，その中でも提案法の RMCM2-CF の AUC が最も高いことが確認できる．また，RMCM-CF および RMCM2-CF は C に関わらず高い AUC を維持していることがわかる．MovieLens-100k データの場合，全体的に item-based 手法の方が user-based 手法より高い値を示した．

表 1: NEEDS-SCAN/PANEL データにおける目的関数値と AUC の相関係数

	user-based	**item-based**
$C = 3, \tau = 10$	-0.8072	-0.6567
$C = 3, \tau = 30$	-0.3583	-0.0579
$C = 5, \tau = 10$	-0.2134	-0.6595

表 2: MovieLens-100k データにおける目的関数値と AUC の相関係数

	user-based	**item-based**
$C = 3, \tau = 10$	-0.7782	-0.8027
$C = 3, \tau = 30$	-0.9763	-0.8980
$C = 5, \tau = 10$	-0.5911	-0.7493

次に，同一パラメータにおける多試行での目的関数値と AUC の間の相関係数を表 1 および 2 に示す．表 1 から，NEEDS-SCAN/PANEL データの場合，$C = 3$, $\tau = 30$ の試行を除くと，全体的に負の相関があることが確認できる．また，表 2 から，MovieLens-100k データの場合 C や τ を変更しても -1 に近い相関関係を維持していることが確認できる．したがって，目的関数値が小さいほど推薦性能が良いという傾向が見られ，RMCM2 法の目的関数設定の妥当性が示唆された．

4. おわりに

本研究では，目的関数ベースの RMCM 法に基づく CF をユーザベースとアイテムの 2 つのアプローチに基づいて提案し，実データに適用することで推薦性能の変化を観察した．実験結果から，RMCM2-CF は RMCM-CF と同様，HCM-CF より高い推薦性能を持つことが確認でき，MovieLens-100k データのように十分なアイテムの数がある場合は item-based 手法が user-based 手法より有効であることが確認できた．また，目的関数値と推薦性能に負の相関があることを確認できた．これより，ラフ集合理論に基づく不確実性の取り扱いが CF タスクにおいて有効であり，かつ RMCM2 法の目的関数の設定には妥当性があることが示唆された．また，提案法において目的関数を改良することでさらに有効な CF 手法を提案することも期待できる．今後の課題としては，欠測値を処理するための機構の導入が挙げられる．

参考文献

[1] H. Kim, S. Ubukata, A. Notsu, and K. Honda: Two types of collaborative filtering membership C-means clustering, Proceeding of The 22nd International Symposium on Advanced Intelligent Systems, 118/123 (2021)

[2] S. Ubukata, A. Notsu, and K. Honda: Objective function-based rough membership C-means clustering, Information Sciences, **548**, 479/496 (2021)

[3] J. Breese, D. Heckerman, and C. Kadie: Empirical analysis of predictive algorithms for collaborative filtering, Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 43/52 (1998)

[4] G. Linden, B. Smith, and J. York: Amazon. com recommendations: Item-to-item collaborative filtering, IEEE Internet computing, **7**-11, 76/80 (2003)