卒業研究論文

題　　目

Rough Co-clustering Induced by Multinomial Mixture Models 法に基づく協調フィルタリングに関する研究

A Study on Collaborative Filtering Based on Rough Co-clustering Induced by Multinomial Mixture Models

研究グループ　　人間情報システム研究室

指導教員　生方誠希准教授，本多克宏教授

令和 3 年 （ 2021 年 ）度卒業

(No. 1181201161)　　毛利 憲竜

大阪府立大学工学域電気電子系学類情報工学課程

1 はじめに

協調フィルタリング (Collaborative Filtering, CF) は，各ユーザーに対し，他のユーザーの趣味嗜好に基づいて好ましいコンテンツの推薦を行う手法であり，Amazon などの電子商取引サイトや YouTube などの動画配信サイト等のコンテンツ推薦システムで広く活用されている．協調フィルタリングにおいては，アイテムベース協調フィルタリング [1] やベイジアン協調フィルタリング [2] など様々な手法が提案されているが，その中でも，クラスタリングベースの協調フィルタリングは実装が容易であり，効率的に計算ができることに加えて，メモリ消費量を低減できるという利点を持っている．クラスタリングとは，データ内の類似した対象をクラスターとして抽出し，自動的にグルーピングを行う手法である．

代表的なクラスタリング手法として，Hard C-Means (HCM; k-Means) 法 [3] がある．HCM 法では，各対象は唯一のクラスターに帰属するよう，排他的な分割が行われる．しかし，協調フィルタリングが対象とするユーザーの嗜好情報は人間の主観的な評価に基づいており，曖昧性や不確実性を含んでいる．したがって，ラフ集合理論 [4] に基づいて不確実性を取り扱うことのできるクラスタリング手法であるラフクラスタリング [5] が有効であると考えられる．ラフクラスタリングは，対象のクラスターに対する帰属の確実性・可能性・不確実性を考慮することにより，一つの対象の複数のクラスターに対する帰属を表現でき，クラスターのオーバーラップを取り扱える．ラフクラスタリングのアルゴリズムとしては，Generalized Rough C-Means (GRCM) 法 [6]，Rough Set C-Means (RSCM) 法 [7]，Rough Membership C-Means (RMCM) 法 [8] など，様々な手法が提案されている．本論文では，GRCM 法において正規化メンバシップを用いてクラスター中心を算出する GRCM with Membership Normalization (GRCM-MN) 法 [8] を採用し，単に RCM 法とよぶ．種々のラフクラスタリング手法をベースにした協調フィルタリングとして，RCM-CF[9]，RSCM-CF[10]，RMCM-CF[11] が提案されている．

また，文書におけるキーワードの頻度，ユーザーの購買履歴などの対象と項目間の共起情報を表す共起関係データのクラスタリングにおいて，関連性の強い対象と項目の組からなる共クラスターを抽出する共クラスタリングが注目されている．協調フィルタリングで扱うユーザーの嗜好情報に関するデータは共起関係データと考えられ，共クラスタリングによる分析が有効であると考えられる．共クラスタリングの手法として Fuzzy Co-Clustering induced by Multinomial Mixture models (FCCMM) 法 [12] が提案されており，FCCMM 法の特殊な場合である Hard CCMM (HCCMM) 法をベースとし，ラフ集合理論の観点を導入したラフ共クラスタリング手法として Rough CCMM (RCCMM) 法 [13] がある．

本研究では，HCCMM 法に基づく協調フィルタリング (HCCMM-CF) と RCM 法に基づく協調フィルタリング (RCM-CF) を参考にして，RCCMM 法に基づく協調フィルタリング (RCCMM-CF) を提案し，実データを用いた数値実験を通して，推薦性能を検証する．また，HCCMM-CF および RCM-CF との比較を通じて提案法の協調フィルタリングタスクにおいての有効性を検証する．

本論文の構成は以下の通りである．第 2 章では，HCM 法，RCM 法，共クラスタリング，HCCMM 法，RCCMM 法および RCM-CF について概説し，第 3 章では，提案法である RCCMM-CF を説明する．第 4 章では，数値実験の設定，結果および各結果の考察を示し，第 6 章で本論文のまとめを述べる．

2 準備

2.1 HCM 法

代表的な非階層的クラスタリング手法である HCM 法は，クラスター中心の算出と対象のクラスター割り当てを交互に繰り返すことでクラスターを抽出する．

HCM 法のアルゴリズムを以下に示す．

Step 1 クラスター数 C を設定する．

Step 2 C 個の初期クラスター中心 $\boldsymbol{b}_c$ を対象空間 $U = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_n\}$ の中から非復元抽出により決定する．

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．

$$d_i^{\min} = \min_{1 \le l \le C} \|\boldsymbol{x}_i - \boldsymbol{b}_l\|, \tag{1}$$

$$u_{ci} = \begin{cases} 1 & \left(d_{ci} \le d_i^{\min}\right), \\ 0 & (\text{otherwise}). \end{cases} \tag{2}$$

Step 4 クラスター中心 $\boldsymbol{b}_c$ を計算する．

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n} u_{ci}\boldsymbol{x}_i}{\sum_{i=1}^{n} u_{ci}}. \tag{3}$$

Step 5 u_{ci} に変化がなくなるまで **Step 3-4** を繰り返す．

HCM 法では，各対象は唯一のクラスターに帰属するため，複数のクラスターへの帰属を表現することができない．よって，HCM 法はデータに内在する曖昧性・不確実性を取り扱うことができない．

2.2 RCM 法

RCM 法は，HCM 法をラフ集合理論によって拡張した手法であり，ラフ集合理論における上近似・下近似・境界領域を模した概念である上エリア・下エリア・境界エリアによって対象がクラスターに属することの可能性・確実性・不確実性を取り扱う．RCM 法では，クラスター割り当てにおいて，オーバーラップ度合いを調節するパラメータを用いて，1 次関数の閾値を増加させることにより，HCM 法の最近隣割り当て ((2) 式) の条件を緩和し，複数の上エリアへの帰属を表現することができる．RCM 法のアルゴリズムを以下に示す．

Step 1 クラスター数 C，オーバーラップ度合いを調節するパラメータ $\alpha \geq 1, \beta \geq 0$ を設定する．

Step 2 C 個の初期クラスター中心 $\boldsymbol{b}_c$ を対象空間 $U = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_n\}$ の中から非復元抽出により決定する．

Step 3 対象 i のクラスター c の上エリアに対するメンバシップ $\overline{u}_{ci}$ と正規化メンバシップ $\tilde{u}_{ci}$ を順に以下の式で求める．

$$\overline{u}_{ci} = \begin{cases} 1 & \left(d_{ci} \leq \alpha d_i^{\min} + \beta\right), \\ 0 & (\text{otherwise}), \end{cases} \tag{4}$$

$$\tilde{u}_{ci} = \frac{\overline{u}_{ci}}{\sum_{l=1}^{C} \overline{u}_{li}}. \tag{5}$$

Step 4 クラスター中心 $\boldsymbol{b}_c$ を計算する．

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^{n} \tilde{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^{n} \tilde{u}_{ci}}. \tag{6}$$

項目 / 対象	項目1	項目2	項目3	項目4	項目5
対象1	1	1	0	0	0
対象2	0	1	1	0	0
対象3	1	1	1	0	0
対象4	0	0	1	1	1
対象5	0	0	0	1	1

共クラスター1（対象1〜3 × 項目1〜3）

共クラスター2（対象4〜5 × 項目4〜5）

図 1　共クラスタリングの概念図

	対象1	対象2	対象3	対象4	対象5
共クラスター1	1	1	1	0	0
共クラスター2	0	0	0	1	1

図 2　対象の共クラスターへの割り当て

	項目1	項目2	項目3	項目4	項目5
共クラスター1	1/3	1/3	1/3	0	0
共クラスター2	0	0	0	1/2	1/2

図 3　項目の共クラスターへの割り当て

Step 5 $\overline{u}_{ci}$ に変化がなくなるまで **Step 3-4** を繰り返す．

GRCM 法は $\alpha = 1, \beta = 0$ の時，HCM 法と同様の結果を与える．α, β を大きくすると対象は複数の上エリアに帰属しやすくなり，クラスターのオーバーラップが大きくなる．

2.3 共クラスタリング

図 1,2,3 に示すように，共クラスタリングでは，対象と項目の共起関係を表す共起関係データから，親近性の高い対象と項目の組からなる共クラスターを抽出する．

対象 i と項目 j の共起度を r_{ij}，対象 i の共クラスター c に対するメンバシップを u_{ci}，項目 j の共クラスターに対するメンバシップを w_{cj}，対象数を n，項目数を m とし，以下で共クラスタリング手法である HCCMM 法と RCCMM 法について説明する．

2.3.1 HCCMM 法

HCCMM 法は FCCMM 法において，対象の分割に関してハードであり，項目のメンバシップ値のファジィ度を考慮しない特殊なモデルである．HCCMM 法の最適化問題は以下で与えられる．

$$\text{max. } J_{\text{HCCMM}} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ci} r_{ij} \log w_{cj}, \tag{7}$$

$$\text{s.t. } u_{ci} \in \{0,1\}, w_{cj} \in (0,1], \forall c,i,j, \tag{8}$$

$$\sum_{c=1}^{C} u_{ci} = 1, \forall i, \ \sum_{j=1}^{m} w_{cj} = 1, \forall c. \tag{9}$$

また，クラスター c との対象 i の類似度 s_{ci} を次式で定める．

$$s_{ci} = \sum_{j=1}^{m} r_{ij} \log w_{cj}. \tag{10}$$

ここで，$s_{ci} \leq 0$ となる点に注意する．s_{ci} が大きいほど類似度が大きい．HCCMM 法のアルゴリズムを以下に示す.

Step 1 クラスター数 C を設定する.

Step 2 項目メンバシップ w_{cj} を次のように初期化する．ランダムに C 個の対象をサンプリングし，総和が 1 となるように正規化したものを設定する.

$$w_{cj} = \frac{r_{cj}}{\sum_{l=1}^{m} r_{cl}}. \tag{11}$$

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，最も類似度の大きいクラスターとの類似度 $s_i^{\max}$ に基づいて計算する.

$$s_i^{\max} = \max_{1\leq c\leq C} s_{ci}, \tag{12}$$

$$u_{ci} = \begin{cases} 1 & (s_{ci} \geq s_i^{\max}), \\ 0 & (\text{otherwise}). \end{cases} \tag{13}$$

Step 4 項目メンバシップ w_{cj} を更新する.

$$w_{cj} = \frac{\sum_{i=1}^{n} u_{ci} r_{ij}}{\sum_{l=1}^{m}\sum_{i=1}^{n} u_{ci} r_{il}}. \tag{14}$$

Step 5 u_{ci} に変化がなくなるまで **Step 3-4** を繰り返す.

HCM 法と同様に各対象は唯一のクラスターに帰属する.

2.3.2 RCCMM 法

RCCMM 法は HCM 法をラフ集合理論によって拡張した RCM 法と同様に HCCMM 法にラフ集合理論の観点を導入したラフ共クラスタリング手法である．RCCMM 法では，クラスター割り当てにおいて，オーバーラップ度合いを調節するパラメータを用いて，1 次関数の閾値を減少させることにより，HCCMM 法の割り当て ((13) 式) の条件を緩和し，複数の上エリアへの帰属を表現することができる．本研究では，GRCM-MN 法と同様に正規化メンバシップを計算する RCCMM-MN 法を採用し，混乱のない限り，RCCMM-MN 法を単に RCCMM 法と書く．RCCMM 法のアルゴリズムを以下に示す.

Step 1 クラスター数 C，オーバーラップ度合いを調節するパラメータ $\alpha \geq 1, \beta \leq 0$，を設定する.

Step 2 項目メンバシップ w_{cj} を (11) 式によって初期化する.

Step 3 対象 i と最も類似したクラスターとの類似度 $s_i^{\max}$ に基づいて，対象 i のクラスター c の上エリアに対するメンバシップ $\overline{u}_{ci}$ を以下の式で計算し，(5) 式のように正規化メンバシップを求める.

$$\overline{u}_{ci} = \begin{cases} 1 & (s_{ci} \geq \alpha s_i^{\max} + \beta), \\ 0 & (\text{otherwise}), \end{cases} \tag{15}$$

Step 4 項目メンバシップ w_{cj} を更新する.

$$w_{cj} = \frac{\sum_{i=1}^{n} \tilde{u}_{ci} r_{ij}}{\sum_{l=1}^{m}\sum_{i=1}^{n} \tilde{u}_{ci} r_{il}}. \tag{16}$$

Step 5 $\overline{u}_{ci}$ に変化がなくなるまで **Step 3-4** を繰り返す.

RCCMM 法は，$\alpha = 1$，$\beta = 0$ のとき，HCCMM 法と同様の結果を与える．s_{ci} が負の値を取るため，α が大きな値を取るほど，また β が小さい値を取るほど対象は複数の上エリアへ帰属しやすくなり，クラスターのオーバーラップが大きくなる.

2.4 RCM-CF

RCM-CF は，RCM 法によって嗜好の類似したユーザーのクラスターを抽出し，クラスター内で嗜

好度の高いコンテンツを推薦する手法である．RCM-CF の手順を以下に示す．

Step 1 $n \times m$ の評価値行列 $R = \{r_{ij}\}$ をベクトルデータ $X = \{x_{ij}\}$ とみなして RCM 法を適用し，正規化メンバシップ $\tilde{u}_{ci}$ とクラスター中心 $\boldsymbol{b}_c$ を求める．ここで n はユーザー数，m はアイテム数である．

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

$$\hat{r}_{ij} = \sum_{c=1}^{C} \tilde{u}_{ci} b_{cj}. \tag{17}$$

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．

$$\tilde{r}_{ij} = \begin{cases} 1 & (\hat{r}_{ij} \geq \eta)\,, \\ 0 & (\text{otherwise})\,. \end{cases} \tag{18}$$

3　提案法：RCCMM-CF

本研究では，ラフ共クラスタリングに基づく協調フィルタリングとして，RCCMM-CF を提案する．RCCMM-CF は，評価値行列 R に RCCMM 法を適用することで，共起関係データに内在する，人間の感性に起因する不確実性を考慮しながら，嗜好の類似したユーザーのクラスターを抽出し，クラスター内で嗜好度の高いコンテンツを推薦する手法である．RCCMM-CF の手順を以下に示す．

Step 1 $n \times m$ の評価値行列 $R = \{r_{ij}\}$ に対して RCCMM 法を適用し，正規化ユーザーメンバシップ $\tilde{u}_{ci}$ とアイテムメンバシップ w_{cj} を求める．

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

$$\hat{r}_{ij} = \sum_{c=1}^{C} \tilde{u}_{ci} w_{cj}. \tag{19}$$

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 (18) 式のように推薦する．

4　数値実験

実データ (NEEDS-SCAN/PANEL データおよび MovieLens-100k データ) に対して提案法を適用し，オーバーラップ度合いを調節するパラメータ α，β やクラスター数 C による推薦性能の変化を検証した．推薦性能の評価指標には ROC-AUC 指標を用いた．

4.1　実験データ

4.1.1　NEEDS-SCAN/PANEL

NEEDS-SCAN/PANEL データは日本経済新聞社が収集した，2000 年の調査対象の 996 世帯 (ユーザー) の 18 個の製品 (アイテム) に対しての所有有無を表すデータである．評価値 r_{ij} はユーザー i がアイテム j を所有している場合 1，所有していない場合 0 となる．このデータの中でランダムに選んだ 1000 個をテストデータとし，テストデータに対する評価値を未評価値として 0 に置き換えたデータをトレーニングデータとして，実験を行った．
【調査対象の 18 製品（括弧内は所有世帯数)】
自動車 (825)，ピアノ (340)，VTR(933)，ルームエアコン (911)，パソコン (588)，ワープロ (506)，CD(844)，VD(325)，自動二輪車 (294)，自転車 (893)，大型電気冷蔵庫 (858)，中・小電気冷蔵庫 (206)，電子レンジ (962)，オーブン (347)，コーヒーメーカー (617)，電気洗濯機 (986)，衣料乾燥機 (226)，電気乾燥機 (242)

4.1.2　MovieLens-100k

MovieLens-100k データは GroupLens Research (https://grouplens.org/) が収集した，943 人のユーザーが 1,682 個の映画に対して行った 100,000 個の評価値のデータである．このデータは未評価値が多い疎なデータであるため，クラスタリングおよび評価のため，本実験ではこのうち 30 個以上の映画を評価した $n = 690$ のユーザーと，50 人以上のユーザーが評価した $m = 583$ の映画を抽出し，77,201 個の評価を含むデータを作成して使用した．そのうちの約 10%である 7,720 個の評価をテストデータとし，テストデータに対する評価値を未評価値の値に置き換えたデータをトレーニングデータとした．元の評価値は 1 ～ 5 の 5 段階評価である．元の評価値

が 4 以上であれば $r_{ij} = 1$，3 以下であれば $r_{ij} = 0$ に置き換え，未評価値は $r_{ij} = 0.5$ としたデータを用いて，実験を行った．

4.2 評価指標

推薦性能は ROC-AUC 指標によって評価した．ROC (Receiver Operating Characteristic) 曲線は，種々の閾値での偽陽性率 (False Positive Rate, FPR) に対する真陽性率 (True Positive Rate, TPR) をプロットすることで得られ，AUC (Area Under the ROC Curve) は ROC 曲線の下部の面積である．AUC が 0.5 のときランダムな推薦であり，1 に近いほど推薦性能が良いといえる．

4.3 実験結果

4.3.1 NEEDS-SCAN/PANEL

クラスター数 C が $\{1, 2, 3, 5\}$ の時の α による AUC の変化を図 4 に示す．各 AUC の値は β を 0 に固定し，α を $[1.0, 2.0]$ の範囲で，0.1 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均であり，$\alpha = 1.0$ の時，HCCMM-CF と同様の結果を示す．クラスター数 C が 1 の時は各製品の所有の有無の平均値が全ての世帯における推薦度となり，AUC は 0.8416 である．図 4 から α の増加にともなって，AUC が $C = 5$ の時最大で 0.025 ほど高くなり，その後ほぼ一定で推移していることがわかる．また，$C = 2$ の時は $\alpha = 1.4$，$C = 3$ の時は $\alpha = 1.8$，$C = 5$ の時は $\alpha = 1.6$ の時の AUC が最も高くなることが確認できる．

クラスター数 C が $\{1, 2, 3, 5\}$ の時の β による AUC の変化を図 5 に示す．各 AUC の値は α を 1.0 に固定し，β を $[-10, 0]$ の範囲で，1 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均であり，$\beta = 0$ の時，HCCMM-CF と同様の結果を示す．図 4 における $\alpha = 1.0$，図 5 における $\beta = 0$ の時の結果は，パラメータ設定が同じであるが，異なる初期値で実行しているため，最大で $C = 3$ の時，0.011 ほど AUC に違いが生じている．図 5 から β は負数であるため，β の減少にともなって，AUC が $C = 5$ の時最大で 0.030 ほど高くなっていることがわかる．また，

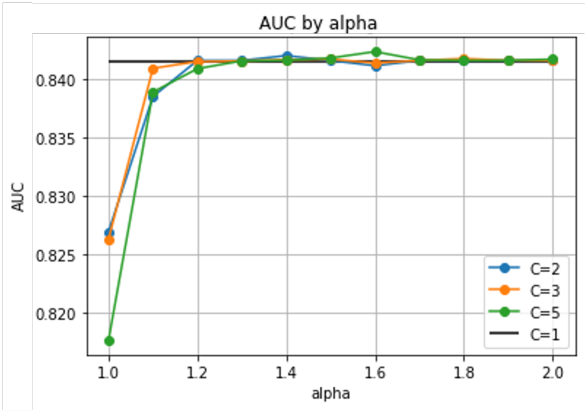

図 4　NEEDS-SCAN/PANEL：$C = 1, 2, 3, 5$ の時の α による AUC の変化 (RCCMM-CF)

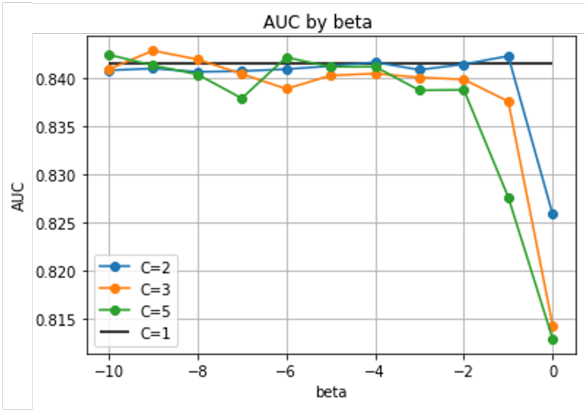

図 5　NEEDS-SCAN/PANEL：$C = 1, 2, 3, 5$ の時の β による AUC の変化 (RCCMM-CF)

$C = 2$ の時は $\beta = -1$，$C = 3$ の時は $\beta = -9$，$C = 5$ の時は $\beta = -10$ の時の AUC が最も高くなることが確認できる．

また，図 4，5 から，α が 1.1 以下，β が -1 以上の場合では大きい推薦性能向上が見られず，それ以降では推薦性能が $C = 1$ の時を超えるなど，高い値でほぼ一定に推移していることがわかる．NEEDS-SCAN/PANEL のデータは項目数が 18 と少なく，製品も自動車，パソコン，冷蔵庫など代表的なものが多いため，クラスター数を増加させると主要な製品との関連が深い世帯がいずれかのクラスターにのみ帰属し，AUC が低下してしまうことが考えられるが，α を増加，β を減少させることで主要な製品との関連が深い世帯のオーバーラップが表現され，AUC を向上させることができていると考えられる．そして，主要な製品が多くの割合を占めていることから，α をより大きく，β をより小さくしても AUC が低下しないと考えられる．

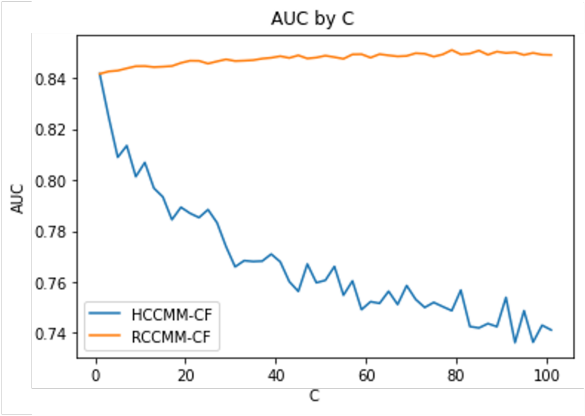

図 6　NEEDS-SCAN/PANEL：クラスター数による AUC の変化 (HCCMM-CF と RCCMM-CF 比較)

クラスター数 C を 1 から 100 まで変化させた時の AUC の変化を図 6 に示す．RCCMM-CF の AUC は，α は $[1.0, 1.5]$ の範囲で 0.05 刻み，β は $[-5, 0]$ の範囲で 0.5 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均値から最大値を採用した．図 6 から，クラスター数に関わらず，提案法が HCCMM-CF より高い AUC 値を持つことが確認できる．また，クラスター数を大きくすると HCCMM-CF の推薦性能は低下する一方で，提案法の場合は安定して推薦性能が向上していることがわかる．

4.3.2　MovieLens-100k

クラスター数 C が $\{1, 2, 3, 5\}$ の時の α による AUC の変化を図 7 に示す．各 AUC の値は β を 0 に固定し，α を [1.000,1.001] の範囲で，0.0001 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均であり，$\alpha = 1.0$ の時，HCCMM-CF と同様の結果を示す．クラスター数 C が 1 の時は各映画の評価値の平均値が全てのユーザーに対する推薦度となり，AUC は 0.6940 である．図 7 から α の増加にともなって，AUC が $C = 5$ の時最大で 0.0119 ほど高くなっていることがわかる．また，$C = 2, 3, 5$ 全てにおいて $\alpha = 1.0005$ の時の AUC が最も高くなり，それ以上の α では低下していることが確認できる．

クラスター数 C が $\{1, 2, 3, 5\}$ の時の β による AUC の変化を図 8 に示す．各 AUC の値は α を 1.0 に固定し，β を $[-3.0, 0]$ の範囲で，0.2 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した

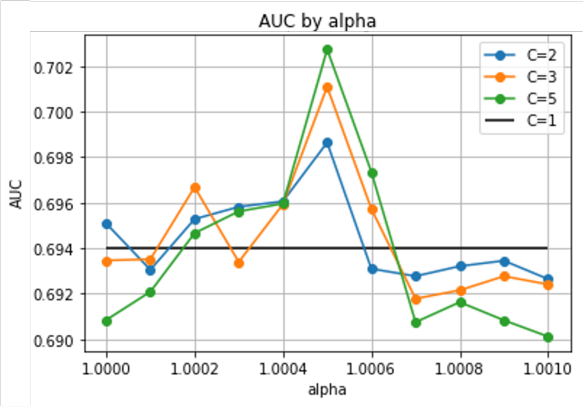

図 7　MovieLens-100k：$C = 1, 2, 3, 5$ の時の α による AUC の変化 (RCCMM-CF)

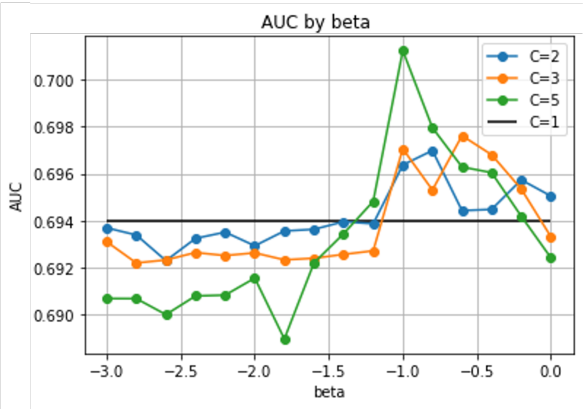

図 8　MovieLens-100k：$C = 1, 2, 3, 5$ の時の β による AUC の変化 (RCCMM-CF)

時のそれぞれの 10 回試行の平均であり，$\beta = 0$ の時，HCCMM-CF と同様の結果を示す．図 7 における $\alpha = 1.0$，図 8 における $\beta = 0$ の時の結果は，パラメータ設定が同じであるが，異なる初期値で実行しているため，最大で $C = 3$ の時，0.004 ほど AUC に違いが生じている．図 8 から β の減少にともなって，AUC が $C = 5$ の時最大で 0.0088 ほど高くなっていることがわかる．また，$C = 2$ の時は $\beta = -0.8$，$C = 3$ の時は $\beta = -0.6$，$C = 5$ の時は $\beta = -1.0$ の時の AUC が最も高くなり，より β が減少すると AUC は低下し，$\beta = -1.4$ 以下ではそれぞれほぼ一定で推移していることがわかる．

α による AUC の変化 (図 7) から，α が 1.0004 ～ 1.0006 の時に AUC が最大となり，それ以上の α では低下していることがわかる．これは，α を 1.0004 ～ 1.0006 まで大きくすると，オーバーラップが表現され，AUC が向上したと考えられるが，それ以上になると，データの項目数が比較的多いため，適切なオー

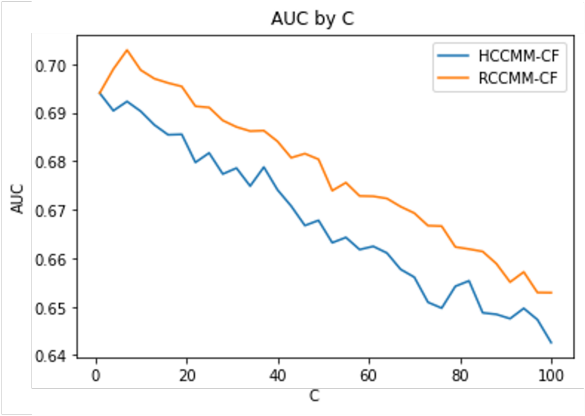

図 9　MovieLens-100k：クラスター数による AUC の変化 (HCCMM-CF と RCCMM-CF の比較)

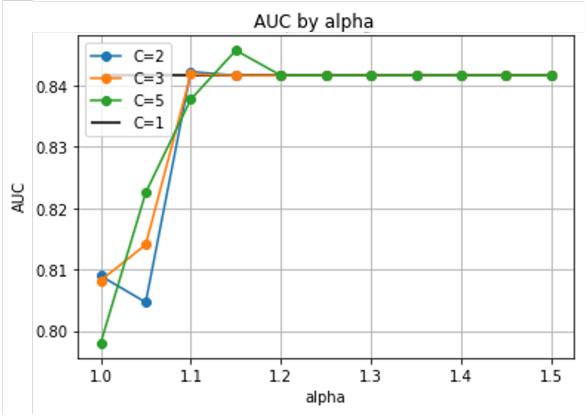

図 10　NEEDS-SCAN/PANEL:$C = 1, 2, 3, 5$ の時の α による AUC の変化 (RCM-CF)

バーラップを超えて，嗜好パターンと関連性の低い項目についても考慮してしまったことで，AUC が低下したと考えられる．また，NEEDS-SCAN/PANEL データに比べ，AUC が最大値となる α が非常に小さい値となっている．これは，用いた評価値行列は未評価値が 82%に至る疎なデータであるため，主要なオーバーラップに対して，未評価値の影響が大きく出てしまうため，推薦性能が向上する α の値が小さくなっていると考えられる．β による AUC の変化 (図 8) でも β が -1 の時に AUC が最大となり，-1 以下では推薦性能が低下していることがわかり，同様のことが考えられる．また，α は類似度との乗算であるのに対し，β は加算となるので，AUC が最大となる β の絶対値は α より大きくなっていると考えられる．

クラスター数 C を 1 から 100 まで変化させた時の AUC の変化を図 9 に示す．RCCMM-CF の AUC は，α は $[1.0000, 1.0005]$ の範囲で 0.0001 刻み，β は $[-2, 0]$ の範囲で 0.2 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均値から最大値を採用した．図 9 から，クラスター数に関わらず，提案法が HCCMM-CF より高い AUC 値を持つことがわかる．また，クラスター数の増加とともに HCCMM-CF の AUC は低下し，RCCMM-CF では $C = 7$ で最大値 (0.7029) となった後低下していることが確認できる．

NEEDS-SCAN/PANEL データ，MovieLens-100k データでのベースとなる HCCMM-CF との比較の実験結果 (図 6，9) から，提案法はいずれもクラスター数に関わらず HCCMM-CF より高い推薦性能を持つことが確認され，ラフ集合理論に基づいて，HCCMM-CF のハードな割り当ての条件を緩和し，オーバーラップを表現することは，実データにおいて有効であると考えられる．また，図 9 より，クラスター数が 7 以上に大きくなると，類似性が比較的高いと考えられる対象群に対して，クラスター数が多くなり，過度に分割を行ってしまうことで，個々のクラスター内での情報量が少なくなり，AUC が低下したと考えられる．

4.3.3　RCM-CF との比較

NEEDS-SCAN/PANEL のデータに対して，RCM-CF を適用した結果は次のようになる．$C = 2, 3, 5$ の時の RCM-CF の α による AUC の変化を図 10 に示す．各 AUC の値は α を $[1.0, 1.5]$ の範囲で，0.05 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均である．図 10 から $C = 2, 3$ の時は $\alpha = 1.1$，$C = 5$ の時は $\alpha = 1.15$ の時の AUC が最も高くなり，α が 1.2 以上では一定で推移していることが確認できる．

NEEDS-SCAN/PANEL データにおいて提案法および RCM-CF のクラスター数を 1 から 100 まで変化させた時の AUC の変化を図 11 に示す．各 AUC は RCM-CF では α を $[1.0, 1.4]$ の範囲で 0.05 刻み，β を $[0, 0.3]$ の範囲で 0.05 刻み，RCCMM-CF では α を $[1.0, 1.6]$ の範囲で 0.05 刻み，β を $[-10, 0]$ の範囲で 1 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均値から最大値を採用した．図 11 より，RCM-CF はクラスター数の増加に伴い AUC の値は高くなっていき，$C = 5$ で最大値 (AUC=0.8488) となり，その後低下していくのに対し，RCCMM-CF はクラス

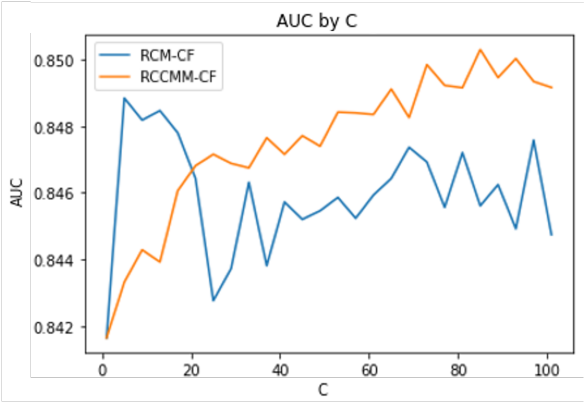

図 11　NEEDS-SCAN/PANEL: クラスター数による AUC の変化 (RCM-CF と RCCMM-CF の比較)

ター数の増加とともに AUC が向上し続けていることが確認できる．RCM-CF はクラスター内の対象の平均であるクラスター中心を推薦に用いるため，クラスター分割が多くなると，クラスター内で相対的に主要な対象の影響が小さくなり，推薦性能が低下したと考えられるが，提案法は項目メンバシップとして共クラスター内での項目の重要度を保持するので，クラスター分割が多くなっても主要な項目の情報が失われることなく，推薦性能を向上させられていると考えられる．

MovieLens-100k のデータに対して，RCM-CF を適用した結果は次のようになる．$C = 2, 3, 5$ の時の RCM-CF の α による AUC の変化を図 12 に示す．各 AUC の値は α を $[1.0, 1.1]$ の範囲で，0.01 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均である．図 12 から $C = 2$ の時は $\alpha = 1.02$，$C = 3, 5$ の時は $\alpha = 1.03$ の時の AUC が最も高くなり，α が $1.03 \sim 1.06$ では低下し，1.06 以上では一定で推移していることが確認できる．

MovieLens-100k のデータにおいて提案法および RCM-CF のクラスター数を 1 から 100 まで変化させた時の AUC の変化を図 13 に示す．各 AUC はどちらも $\beta = 0$ に固定し，RCM-CF では α を $[1.0, 1.1]$ の範囲で 0.01 刻み，RCCMM-CF で α を $[1.000, 1.001]$ の範囲で 0.0001 刻みで変化させ，各試行ごとに初期値をランダムに設定して実行した時のそれぞれの 10 回試行の平均値から最大値を採用した．図 13 より，RCM-CF の AUC はクラスター数の増加に伴い AUC の値は高くなっていき，$C = 16$ で最大値 (AUC=0.7289) となった後低下し，RCCMM-CF の

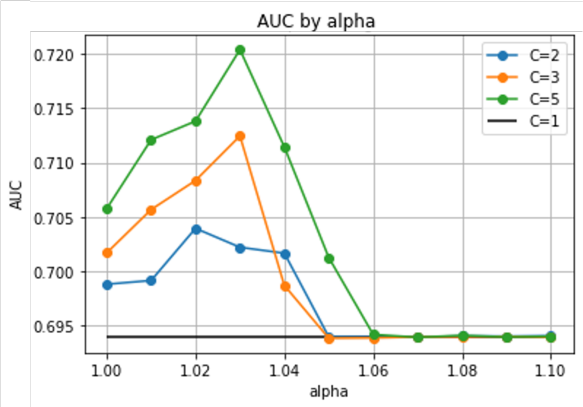

図 12　MovieLens-100k:$C = 2, 3, 5$ の時の α による AUC の変化 (RCM-CF)

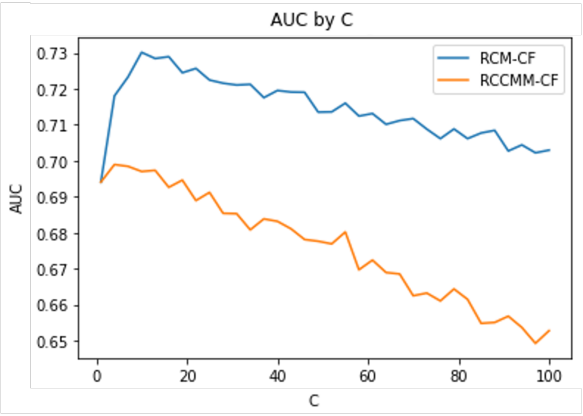

図 13　MovieLens-100k: クラスター数による AUC の変化 (RCM-CF と RCCMM-CF の比較)

AUC は $C = 4$ で最大値 (AUC=0.6989) となった後低下することが確認できる．また，クラスター数に関わらず，提案法の AUC は RCM-CF に劣ることが確認できる．RCM-CF はクラスター内のユーザーの平均評価値であるクラスター中心をもとに推薦度を計算するのに対し，RCCMM-CF ではクラスター内での項目の重要度を表す項目メンバシップをもとに推薦度を計算するため，データの項目の特徴の影響を RCM-CF よりも受けやすく，MovieLens-100k データは項目数が多く，未評価値が占める割合が多いデータであったため，未評価値の影響を強く受けることで推薦性能が向上せず，RCM-CF より劣る結果となったと考えられる．

5 おわりに

本研究では，RCCMM 法に基づく協調フィルタリングを提案し，実データである NEEDS-SCAN/PANEL データおよび MovieLens-100k データに適用し，推薦性能を検証した．実験結果から，NEEDS-SCAN/PANEL データにおいてパラメータを適切に設定することで，提案法が HCCMM-CF より高い性能を持つことが確認できた．MovieLens-100k データにおいても，適切なパラメータ設定により，HCCMM-CF より高い推薦性能を持つことが確認できた．これにより，ラフ集合理論に基づく不確実性の取り扱いが協調フィルタリングタスクにおいて有効であることが示唆された．また，NEEDS-SCAN/PANEL データにおける提案法と RCM-CF の比較において，提案法の推薦性能が RCM-CF を上回ったが，MovieLens-100k データにおける RCM-CF と提案法の比較においては，提案法の推薦性能は RCM-CF に劣ることが確認された．今後の課題としては，欠測値を適切に処理する機構の導入や，パラメータ設定の基準の提案などが挙げられる．

謝辞

本研究は本学工学域電気電子系学類情報工学課程の生方誠希准教授，本多克宏教授，現代システム科学域環境システム学類人間環境科学課程の野津亮教授の御指導のもとに行われたものであり，心より感謝の意を表します．

参考文献

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl: Item-based Collaborative Filtering Recommendation Algorithms, Proceedings of the 10th International Conference on World Wide Web, 285-295 (2001)

[2] X. Su and T. M. Khoshgoftaar: A Survey of Collaborative Filtering Techniques, Advances in Artificial Intelligence, 2009, #421425, 1-19 (2009)

[3] J. MacQueen: Some Methods of Classification and Analysis of Multivariate Observations, Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, 281-297 (1967)

[4] Z. Pawlak: Rough Sets, International Journal of Computer & Information Sciences, 11, 5, 341-356 (1982)

[5] 生方 誠希: ラフ集合に基づく C-Means 型クラスタリングの展開, 日本知能情報ファジィ学会誌, 32, 4, 121-127 (2020)

[6] S. Ubukata, A. Notsu, and K. Honda: General Formulation or Rough C-means Clustering, International Journal of Coumputer Sceince and Network Security, 17, 9, 29-38 (2017)

[7] S. Ubukata, K. Umado, A. Notsu, and K. Honda: Characteristics of Rough Set C-Means Clustering, Journal of Advanced Computational Intelligence and Intelligent Informatics, 22, 4, 551-564 (2018)

[8] S. Ubukata, H. Kato, A. Notsu, and K. Honda: Rough Set-based Clustering Utilizing Probabilistic Membership, Journal of Advanced Computational Intelligence and Intelligent Informatics, 22, 6, 956-964 (2018)

[9] S. Ubukata, S. Takahashi, A. Notsu, and K. Honda: Basic Consideration of Collaborative Filtering Based on Rough C-Means Clustering, Proceedings of Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems, 256-261 (2020)

[10] S. Ubukata, Y. Murakami, A. Notsu, and K. Honda: Basic Consideration of Collaborative Filtering Based on Rough Set C-means Clustering, Proceedings of 22nd International Symposium on Advanced Intelligent Systems, #OS19-4 (2021)

[11] キム ヘラン, 生方 誠希, 本多 克宏, 野津 亮: Rough Membership C-Means 法に基づく協調フィルタリングに関する一考察, 計測自動制御学会関西支部・システム制御情報学会シンポジウム講演論文集, #B1-2, 33-34 (2021)

[12] K. Honda, S. Oshio, and A. Notsu: Fuzzy Co-clustering Induced by Multinomial Mixture Models, Journal of Advanced Computational Intelligence and Intelligent Informatics, 19, 6, 717-726 (2015)

[13] S. Ubukata, N. Nodake, A Notsu, and K. Honda: Basic Consideration of Co-clustering Based on Rough Set Theory, Proceedings of 8th International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, 151-161 (2020)