粒状性を考慮したラフ集合ベースの混合多項分布型共クラスタリングに基づく協調フィルタリング

Collaborative Filtering Based on Rough Set-Based Co-clustering Induced by Multinomial Mixture Models Considering Uncertainty

大阪公立大学大学院　情報学研究科　基幹情報学専攻
知能情報工学分野　人間情報システム研究グループ　BGA22050　毛利 憲竜

Abstract : In clustering-based collaborative filtering (CF), clusters of users with similar preference patterns are extracted, and items with high preferences within the cluster are recommended. Since data in CF tasks contain uncertainties arising from human sensibilities, represented as co-occurrence relationships between users and items, approaches such as rough clustering and co-clustering can be effective. Thus, rough co- clustering induced by multinomial mixture models (RCCMM) and its application to CF (RCCMM-CF) have been proposed. However, RCCMM has a problem in that it does not consider the granularity, an important viewpoint in rough set theory. In this study, we propose a CF approach based on rough set-based co-clustering induced by multinomial mixture models (RSCCMM) considering granularity. Furthermore, we verify the recommendation performance of the proposed method through numerical experiments using real-world datasets.

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．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が，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，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ラフクラスタリング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，Rough Set C-Means (RSCM) 法, Rough Membership C-Means (RMCM) 法など，様々な手法が提案されている [5]．また，これら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が提案されている [6, 7, 8].

そして，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データはユーザー×アイテムの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ファジィ理論に基づく Fuzzy Co-Clustering induced by Multinomial Mixture models (FCCMM) 法 [9] や 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 Rough CCMM (RCCMM) 法 [10] がある．また，RCCMM 法に基づく CF (RCCMM-CF) が提案されている [11]．しかし，RCCMM 法はラフ集合理論において重要な概念である対象空間の粒状性を考慮しておらず，ラフ近似を定義通りに使用していないという問題があるため，Ubukata *et al.* は，対象空間の粒状性を考慮したラフ共クラスタリング手法として Rough Set CCMM (RSCCMM) 法 [12] を提案した．

本研究では，RSCCMM 法に基づき，ユーザー集合の粒状性を考慮することのできる協調フィルタリング (RSCCMM-CF) を提案し，実データを用いた数値実験を通してその推薦性能を検証する．また，

従来の HCM 法に基づく協調フィルタリング (HCM-CF) および RCCMM 法に基づく協調フィルタリング (RCCMM-CF) との比較を通じて提案法の有効性を検証し，ラフ集合理論における粒状化の協調フィルタリングタスクにおける効果について考察を行う．

本論文は以下の 7 章から構成されている．第 2 章では，準備として各クラスタリング手法 (HCM 法，RCM 法，RSCM 法) について概説し，第 3 章では，共クラスタリング手法 (HCCMM 法，RCCMM 法，RSCCMM 法)，RCCMM-CF および提案法である RSCCMM-CF を説明する．第 4 章では数値実験の設定を，第 5 章では結果を示す．最後に，第 6 章で考察を，第 7 章で本論文のまとめを述べる．

2 準備

各クラスタリング手法を概説するにあたり，n 個の対象からなる全体集合 $U = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_i, ..., \boldsymbol{x}_n\}$ から C 個のクラスターを抽出する問題を考える．各対象は m 次元ベクトル $\boldsymbol{x}_i = (x_{i1}, ..., x_{ij}, ..., x_{im})^\top$ で表されるとし，各クラスター c はクラスター中心 $\boldsymbol{b}_c = (b_{c1}, ..., b_{cj}, ..., b_{cm})^\top$ を持つとする．

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．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 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．

$$d_{ci} = \|\boldsymbol{x}_i - \boldsymbol{b}_c\|, \tag{1}$$

$$d_i^{\min} = \min_{1 \le l \le C} d_{li}, \tag{2}$$

$$u_{ci} = \begin{cases} 1 & \left(d_{ci} \le d_i^{\min}\right), \\ 0 & (\text{otherwise}). \end{cases} \tag{3}$$

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

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

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

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 ((3) 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 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 [5] を採用し，単に RCM 法とよぶ．

RCM 法のアルゴリズムを以下に示す．

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

$$\overline{u}_{ci} = \begin{cases} 1 & \left(d_{ci} \le \alpha d_i^{\min} + \beta\right), \\ 0 & (\text{otherwise}), \end{cases} \tag{5}$$

$$\tilde{u}_{ci} = \frac{\overline{u}_{ci}}{\sum_{l=1}^C \overline{u}_{li}}. \tag{6}$$

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

$$\boldsymbol{b}_c = \frac{\sum_{i=1}^n \tilde{u}_{ci}\boldsymbol{x}_i}{\sum_{i=1}^n \tilde{u}_{ci}}. \tag{7}$$

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

2.3 RSCM 法

RSCM 法は粒状性を考慮したラフクラスタリングであり，対象空間 U 上の二項関係 $R \subseteq U \times U$ を用いて，対象空間の粒状化を行う．各対象を二項関係 R による近傍に基づいてクラスター割り当てを行うことで，確実性・可能性・不確実性を取り扱う．この二項関係の設定の仕方によって多様な分類が可能になる．$n \times n$ の行列要素を用いて，対象間の関係の有無を次式で表す:

$$R_{it} = \begin{cases} 1 & (x_i R x_t)\,, \\ 0 & (\text{otherwise})\,. \end{cases} \tag{8}$$

RSCM 法は，RCM 法と同様に明確な目的関数を持たないが，各対象の正規化メンバシップ $\tilde{u}_{ci}$ と各クラスター中心 $\boldsymbol{b}_c$ の交互更新により実行される．

RSCM 法のアルゴリズムを以下に示す．

Step 1 クラスター数 C，二項関係 $R \subseteq U \times U$ 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 の中から非復元抽出により決定する．

Step 3 暫定クラスターに対するメンバシップ u_{ci} を (3) 式で求める．

Step 4 対象 i のクラスター c に対するラフメンバシップ値 μ_{ci}^R と上近似に対するメンバシップ $\overline{u}_{ci}$ を順に以下の式で計算し，正規化メンバシップ値 $\tilde{u}_{ci}$ を (6) 式で計算する．

$$\mu_{ci}^R = \frac{\sum_{t=1}^{n} R_{it} u_{ct}}{\sum_{t=1}^{n} R_{it}} \tag{9}$$

$$\overline{u}_{ci} = \begin{cases} 1 & (\mu_{ci}^R > 0)\,, \\ 0 & (\text{otherwise})\,. \end{cases} \tag{10}$$

Step 5 クラスター中心 $\boldsymbol{b}_c$ を (7) 式で計算する．

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

2.4 共クラスタリング

図 1,2 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．

図 1: 共クラスタリングの概念図

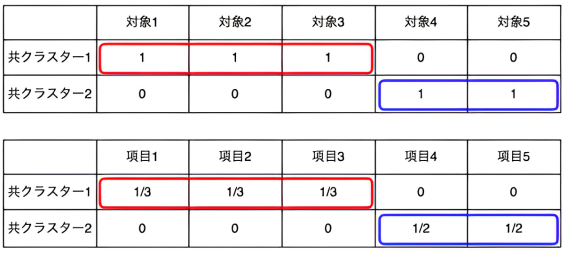

図 2: 対象と項目の共クラスターへの割り当て

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 c に対するメンバシップを w_{cj}，対象数を n，項目数を m とする．以下で共クラスタリング手法である HCCMM 法と RCCMM 法，RSCCMM 法について説明する．

2.4.1 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．

$$\max.\ J_{\text{HCCMM}} = \sum_{c=1}^{C}\sum_{i=1}^{n}\sum_{j=1}^{m} u_{ci} r_{ij} \log w_{cj}, \tag{11}$$

$$\text{s.t.} u_{ci} \in \{0,1\}, w_{cj} \in (0,1], \forall c,i,j, \tag{12}$$

$$\sum_{c=1}^{C} u_{ci} = 1, \forall i,\ \sum_{j=1}^{m} w_{cj} = 1, \forall c. \tag{13}$$

クラスター c と対象 i の類似度 s_{ci} を次式で定める．

$$s_{ci} = \sum_{j=1}^{m} r_{ij} \log w_{cj}. \tag{14}$$

ここで，$s_{ci} \leq 0$ となる点に注意する．s_{ci} が大きいほどクラスター c と対象 i が類似していると判断される．HCCMM 法は HCM 法と同様に各対象は唯一のクラスターに帰属し，対象の分割に関してハードである．

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する．

$$w_{cj} = \frac{r_{cj}}{\sum_{l=1}^{m} r_{cl}}. \tag{15}$$

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．

$$s_i^{\max} = \max_{1 \leq c \leq C} s_{ci}, \tag{16}$$

$$u_{ci} = \begin{cases} 1 & (s_{ci} \geq s_i^{\max}), \\ 0 & (\text{otherwise}). \end{cases} \tag{17}$$

Step 4 項目メンバシップ w_{cj} を更新する．

$$w_{cj} = \frac{\sum_{i=1}^{n} u_{ci} r_{ij}}{\sum_{l=1}^{m} \sum_{i=1}^{n} u_{ci} r_{il}}. \tag{18}$$

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

2.4.2 RCCMM 法

RCCMM 法は 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，クラスターの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 ((17) 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，正規化メンバシップに基づいて項目メンバシップを計算する RCCMM-MN 法を採用し，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, クラスターの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 ランダムに C 個の対象をサンプリングし，項目メンバシップ w_{cj} を (15) 式によって初期化する．

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 $\overline{u}_{ci}$ を以下の式で計算し，(6) 式で正規化メンバシップを求める．

$$\overline{u}_{ci} = \begin{cases} 1 & (s_{ci} \geq \alpha s_i^{\max} + \beta), \\ 0 & (\text{otherwise}). \end{cases} \tag{19}$$

Step 4 項目メンバシップ w_{cj} を更新する．

$$w_{cj} = \frac{\sum_{i=1}^{n} \tilde{u}_{ci} r_{ij}}{\sum_{l=1}^{m} \sum_{i=1}^{n} \tilde{u}_{ci} r_{il}}. \tag{20}$$

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

2.4.3 RSCCMM 法

RSCCMM 法は対象空間の粒状性を考慮したラフ共クラスタリング手法である．対象空間 U が二項関係 $R \subseteq U \times U$ によって粒状化されたと想定し，R の設定の仕方によって多様な分類を可能にする．RSCCMM 法は対象空間の粒状性を考慮したラフ共クラスタリング手法であり，RCCMM 法と同様に HCCMM 法を基礎として，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．本研究では，正規化メンバシップに基づいて項目メンバシップを計算する RSCCMM-MN 法を採用し，単に RSCCMM 法と書く．

RSCCMM 法のアルゴリズムを以下に示す．

Step 1 クラスター数 C，二項関係 $R \subseteq U \times U$ を設定する．

Step 2 ランダムに C 個の対象をサンプリングし，項目メンバシップ w_{cj} を (15) 式によって初期化する．

Step 3 暫定クラスターに対するメンバシップ u_{ci} を (17) 式で求める．

Step 4 対象 i のクラスター c に対するラフメンバシップ値 μ_{ci}^{R} と上近似に対するメンバシップ $\overline{u}_{ci}$, 正規化メンバシップ $\tilde{u}_{ci}$ を (9), (10), (6) 式で順に求める．

Step 5 項目メンバシップ w_{cj} を (20) 式で更新する.

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

2.5 RCCMM-CF

RCCMM-CF[11] は，評価値行列 $X = \{r_{ij}\}$ 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 $X = \{r_{ij}\}$ 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.

$$\hat{r}_{ij} = \sum_{c=1}^{C} \tilde{u}_{ci} w_{cj}. \tag{21}$$

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.

$$\check{r}_{ij} = \begin{cases} 1 & (\hat{r}_{ij} \geq \eta)\,, \\ 0 & (\text{otherwise})\,. \end{cases} \tag{22}$$

3 提案法：RSCCMM-CF

本研究では，対象空間の粒状性を考慮した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，RSCCMM-CF を提案する．RSCCMM-CF は，対象空間の粒状化を通して，ラフ集合理論におけるラフ近似を定義通りに使用して推薦を行う.

3.1 二項関係の設定

対象空間を粒状化するため，対象間の二項関係を設定する．まず，二項関係を構成するための対象間の類似度を定義する．共起関係データに適した類似度を考えるため，HCCMM 法における対象 i とクラスター c の類似度 ((14)) 式を参考に，対象 i と対象 t の類似度 S_{it} を定義する．項目メンバシップは混合多項分布から派生したものであるため，確率分布を基礎とした類似度を考える．そこで，各対象について共起情報の総和が 1 となるように正規化した $\tilde{r}_{ij}$ を考慮し，各対象の共起情報を確率分布として捉える:

$$\tilde{r}_{ij} = \frac{r_{ij}}{\sum_{l=1}^{m} r_{il}}. \tag{23}$$

対象 i と対象 t の類似度 S_{it} を下記のように定義する:

$$S_{it} = \sum_{j=1}^{m} \tilde{r}_{tj} \log \tilde{r}_{ij}. \tag{24}$$

これは，負の交差エントロピーとみなせる．類似度 S_{it} に基づき，二項関係を以下のように設定する:

$$R_{it} = \begin{cases} 1 & (S_{it} \geq \delta)\,, \\ 0 & (\text{otherwise})\,. \end{cases} \tag{25}$$

ここで，$\delta \leq 0$ はラフさを調節するパラメータであり，δ が小さいほど粗い粒状化となり，上近似が拡大し，クラスターのオーバーラップが大きくなる．一般に，S_{it} は非対称であり，R_{it} は対称性を満たさない.

3.2 アルゴリズム

RSCCMM-CF の手順を以下に示す.

Step 1 ラフさを調節するパラメータ $\delta \leq 0$ を設定し，$n \times m$ の評価値行列 $X = \{r_{ij}\}$ に対して，対象間の二項関係を (25) 式によって定める．RSCCMM 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 (21) 式で計算する.

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 (22) 式で推薦する.

4 数値実験

2 種類の実データ (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，クラスターのオーバーラップ度合いを調

節するパラメータ δ 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 種類の製品 (アイテム) に対しての所有の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 1,000 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 本の映画に対して行った 100,000 個の，1 ～ 5 の 5 段階評価値のデータである．このデータに以下の 2 種類の前処理を加え, 二値化を行った MovieLens-100k データセットと二値化を行わない MovieLens-100k データセットを作成し，数値実験を行った．

二値化を行った MovieLens-100k データセットでは，943 人のユーザーと 1,682 本の映画のうち，30 本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 $r_{ij} = 0.5$ とする返還を行った．

二値化を行わない MovieLens-100k データセットでは, データの抽出および二値化を行わず, 943 人のユーザーが 1,682 本の映画に対して行った 100,000 個の, 1 ～ 5 の 5 段階評価値のデータのうちの 10%(10,000 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の値に置き換えたデータをトレーニングデータとした．未評価値に関しては，各ユーザーの平均評価値とした．

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に対する真陽性率をプロットすることで得られ, AUC (Area Under the Curve) は ROC 曲線の下部の面積である．AUC はランダムな推薦のとき 0.5 程度になり，1 に近いほど推薦性能が良いといえる．

5 実験結果

5.1 NEEDS-SCAN/PANEL データセットにおける実験結果

HCCMM-CF，RCCMM-CF および提案法である RSCCMM-CF を NEEDS-SCAN/PANEL データセットに適用した結果を示す．

まず，提案法において種々のクラスター数 $C \in \{1, 3, 5, 7\}$ における δ による AUC の変化を図 3 に示す．各 AUC の値は，$\delta \in [-15.0, -5.0]$ を 1.0 刻みで変化させた時の，異なる初期値による 10 回試行の平均値である．δ は小さいとき粗い粒状化，大きいとき細かい粒状化を表す．$C = 1$ の時は各製品の所有の有無の平均値が全世帯に対する推薦度となり，δ の値に依存せず，AUC は 0.8416 である．図 3 から，$C = \{3, 5, 7\}$ の場合，δ を減少させて粗い粒状化にしていくと，いずれも $\delta = -6.0$ で AUC が最大となり，その後減少し，収束していくことがわかる．また，$C = 1$ の場合の結果より，最大で $C = 3$ の場合は 0.0061，$C = 5$ の場合は 0.0065，$C = 7$ の場合は 0.0072 ほど高い AUC となった．

次に HCCMM-CF，RCCMM-CF および提案法である RSCCMM-CF について，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 4 に示す．各 AUC は，HCCMM-CF は異なる初期値

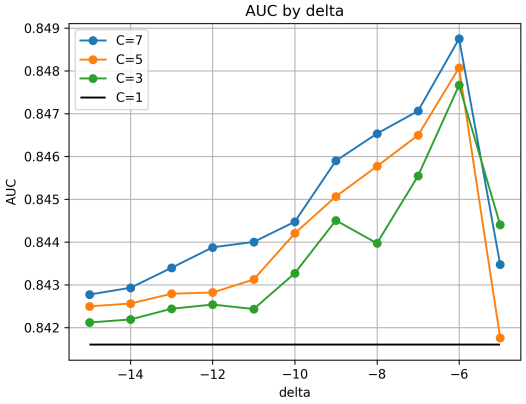

図 3: NEEDS-SCAN/PANEL: 各 C における δ による AUC の変化 (RSCCMM-CF)

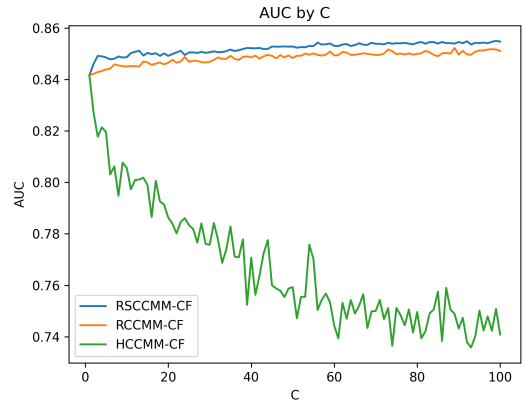

図 4: NEEDS-SCAN/PANEL: C による AUC の変化 (RSCCMM-CF，RCCMM-CF，HCCMM-CF)

による 5 回試行の平均値であり，RSCCMM-CF では $\delta \in [-10.0, -5.0]$ を 0.5 刻み，RCCMM-CF では $\alpha \in [1.2, 1.6]$ を 0.05 刻み，$\beta \in [-9.0, 0.0]$ を 1.0 刻みで変化させ，各々異なる初期値による 5 回試行の平均値から最大値を採用した．図 4 から，クラスター数に関わらず，RCCMM-CF および提案法の RSCCMM-CF が HCCMM-CF より高い AUC を持ち，さらに提案法の RSCCMM-CF が RCCMM-CF より高い AUC を持つことが確認できる．また，クラスター数 C の変化に注目すると，C を大きくすると HCCMM-CF の AUC は落ちる反面，RCCMM-CF および RSCCMM-CF の場合は安定した AUC を持つことも確認できる．

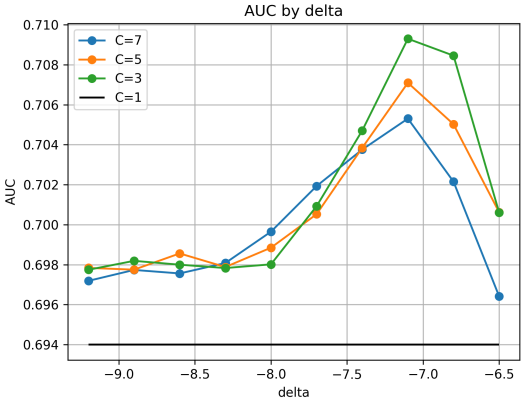

図 5: 二値化を行った MovieLens-100k: 各 C における δ による AUC の変化 (RSCCMM-CF)

5.2 二値化を行った MovieLens-100k データセットにおける実験結果

HCCMM-CF，RCCMM-CF および提案法である RSCCMM-CF を二値化を行った MovieLens-100k データセットに適用した結果を示す．

まず，提案法において種々のクラスター数 $C \in \{1, 3, 5, 7\}$ における δ による AUC の変化を図 5 に示す．各 AUC の値は，$\delta \in [-9.2, -6.5]$ を 0.3 刻みで変化させた時の，異なる初期値による 10 回試行の平均値である．$C = 1$ の時は各映画の評価値の平均値が全ユーザーに対する推薦度となり，δ の値に依存せず，AUC は 0.6940 である．図 5 から，$C = \{3, 5, 7\}$ の場合，δ を減少させて粗い粒状化にしていくと，いずれも $\delta = -7.1$ で AUC が最大となり，その後減少し，収束していくことがわかる．また，$C = 1$ の場合の結果より，最大で $C = 3$ の場合は 0.0153，$C = 5$ の場合は 0.0131，$C = 7$ の場合は 0.0113 ほど高い AUC となった．

次に HCCMM-CF，RCCMM-CF および提案法である RSCCMM-CF について，クラスター数 C を 1 から 30 まで変化させた時の AUC の変化を図 6 に示す．各 AUC は，HCCMM-CF は異なる初期値による 10 回試行の平均値であり，RSCCMM-CF では $\delta \in [-8.1, -6.5]$ を 0.2 刻み，RCCMM-CF では β を 0 に固定し，$\alpha \in [1.0001, 1.001]$ を 0.0001 刻み，各々異なる初期値による 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，RCCMM-CF および提案法の RSCCMM-CF が HCCMM-CF より高い AUC を持ち，さらに提案法の RSCCMM-CF が RCCMM-CF より高い AUC を

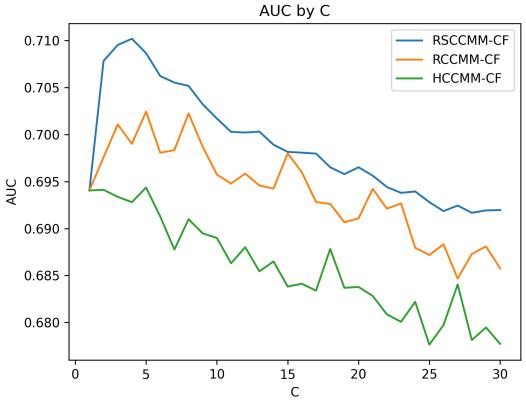

図 6: 二値化を行った MovieLens-100k: C による AUC の変化 (RSCCMM-CF，RCCMM-CF，HCCMM-CF)

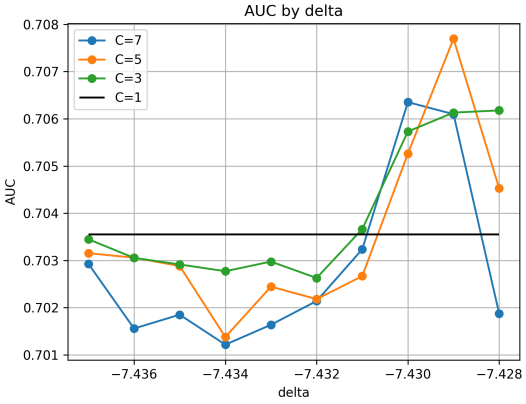

図 7: 二値化を行わない MovieLens-100k: 各 C における δ による AUC の変化 (RSCCMM-CF)

持つことが確認できる．また，クラスター数 C の変化に注目すると，C を大きくするとすべての手法において AUC は向上し，HCCMM-CF，RCCMM-CF では $C = 5$，RSCCMM-CF では $C = 4$ の時に最大となったのち，低下していくことが確認できる．

5.3 二値化を行わない MovieLens-100k データセットにおける実験結果

HCCMM-CF，RCCMM-CF および提案法である RSCCMM-CF を二値化を行わない MovieLens-100k データセットに適用した結果を示す．

まず，提案法において種々のクラスター数 $C \in \{1, 3, 5, 7\}$ における δ による AUC の変化を図 7 に示す．各 AUC の値は，$\delta \in [-7.437, -7.428]$ を 0.001 刻みで変化させた時の，異なる初期値による 10 回試行の平均値である．$C = 1$ の時は各映画の評価値の平均値が全ユーザーに対する推薦度となり，δ の値に依存せず，AUC は 0.7036 である．図 7 から，$C = \{3, 5, 7\}$ の場合，δ を減少させて粗い粒状化にしていくと，$C = 3$ では $\delta = -7.428$ で，$C = 5$ では $\delta = -7.429$ で，$C = 7$ では $\delta = -7.430$ で AUC が最大となり，その後減少し，収束していくことがわかる．また，$C = 1$ の場合の結果より，最大で $C = 3$ の場合は 0.0026，$C = 5$ の場合は 0.0041，$C = 7$ の場合は 0.0028 ほど高い AUC となった．

次に HCCMM-CF および提案法である RSCCMM-CF について，クラスター数 C を 1 から 30 まで変化させた時の AUC の変化を図 8 に示す．各 AUC は，HCCMM-CF では異なる初期値による 5 回試行の平均値から最大値を採用し，RSCCMM-CF では $\delta \in [-7.434, -7.428]$ を 0.001 刻みで各々異なる初期値による 5 回試行の平均値から最大値を採用した．なお，RCCMM-CF では，$\alpha \in [1.0, 1.1]$，$\beta \in [-10, 0]$ において様々なパラメータ設定で実験を行ったが，$C = 1$ の場合より高い AUC を確認することがなかったため，比較を割愛する．図 8 から，クラスター数に関わらず，提案法の RSCCMM-CF が HCCMM-CF より高い AUC を持つことが確認できる．また，クラスター数 C の変化に注目すると，C を大きくすると HCCMM-CF では AUC が低下する反面，RSCCMM-CF では $C = 10$ の時に最大となったのち，低下していくことが確認できる．

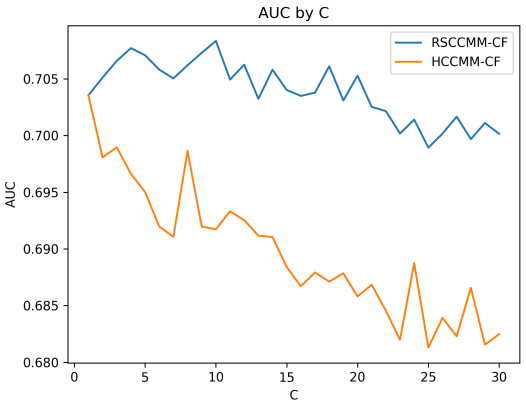

図 8: 二値化を行わない MovieLens-100k: C による AUC の変化 (RSCCMM-CF，HCCMM-CF)

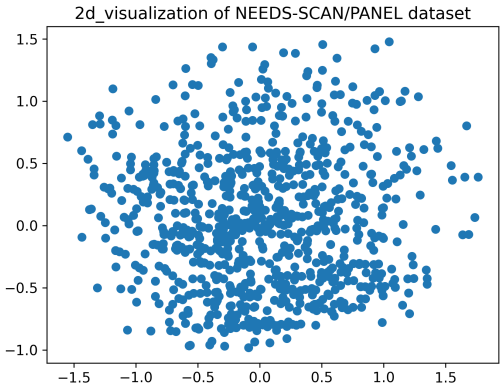

図 9: NEEDS-SCAN/PANEL: PCA による 2 次元への視覚化

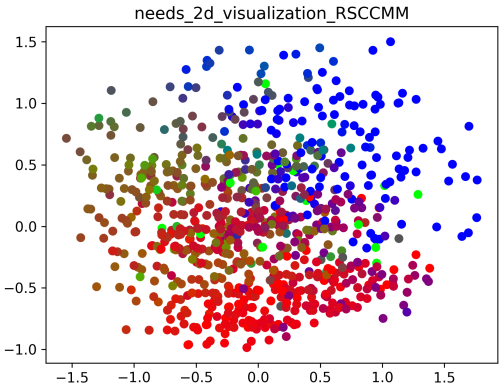

図 10: NEEDS-SCAN/PANEL: PCA による 2 次元への視覚化

5.4 各データセットの視覚化と提案法の適用

NEEDS-SCAN/PANEL, 二値化を行った MovieLens-100k, 二値化を行わない MovieLens-100k の各データセットの評価値行列に RSCCMM 法を適用し，PCA を用いて次元圧縮を行うことで 2 次元に視覚化した結果を示す．

5.4.1 NEEDS-SCAN/PANEL データの視覚化

NEEDS-SCAN/PANEL の評価値行列に PCA を用いて次元圧縮を行うことで 2 次元に視覚化した結果を図 9 に示す．

また，$C = 3$ で，$\delta \in [-8.5, -5.0]$ を 0.5 刻みで変化させ，AUC が最大であった時のラフメンバシップ値によって各データポイントを色分けした結果を図 10 に示す．なお，この時の δ は -6.0 であり，AUC は 0.8515 であった．10 から，データは x 軸の値が 0 より小さく，y 軸の値が 0 より大きい領域のユーザーが多くオーバーラップして，複数のクラスターに帰属しており，その他のデータは主にオーバーラップせず，y 軸の値が 0 より小さい領域と，x 軸の値が 0 より大きく，y 軸の値が 0 より大きい領域に各々分割されている様子が確認できる．

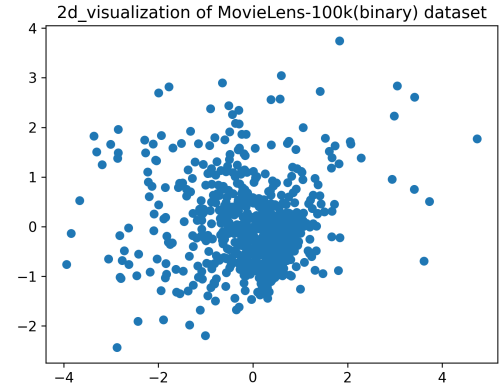

図 11: 二値化を行った MovieLens-100k: PCA による 2 次元への視覚化

5.4.2 二値化を行った MovieLens-100k データの視覚化

二値化を行った MovieLens-100k データの評価値行列に PCA を用いて次元圧縮を行うことで 2 次元に視覚化した結果を図 11 に示す．

また，$C = 3$ で，$\delta \in [-8.9, -6.5]$ を 0.2 刻みで変化させ，AUC が最大であった時のラフメンバシップ値によって各データポイントを色分けした結果を図 12 に示す．なお，この時の δ は -6.9 であり，AUC は 0.7125 であった．12 から，データは x 軸，y 軸の値が 0 付近のユーザーが多くオーバーラップして，複数のクラスターに帰属しており，x 軸，y 軸の値が 0 から離れているユーザーがオーバーラップしておらず，各々唯一のクラスターに分割されている様子が確認できる．

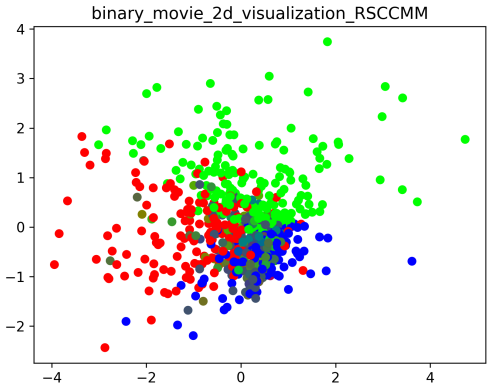

図 12: 二値化を行った MovieLens-100k: PCA による 2 次元への視覚化

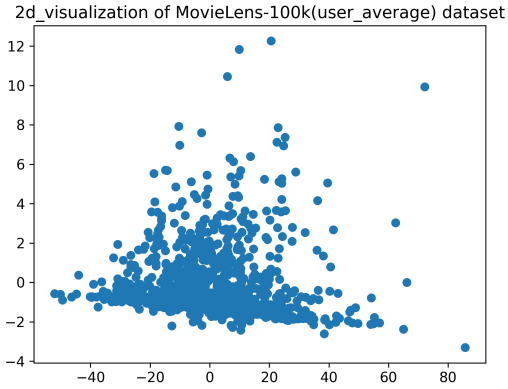

図 13: 二値化を行わない MovieLens-100k: PCA による 2 次元への視覚化

5.4.3 二値化を行わない MovieLens-100k データの視覚化

二値化を行わない MovieLens-100k データの評価値行列に PCA を用いて次元圧縮を行うことで 2 次元に視覚化した結果を図 13 に示す．PCA の結果として得られる第一主成分 (x 軸) はユーザの平均評価値，第二主成分 (y 軸) はユーザが与えた評価値の分散であることが確認できた．

また，$C = 3$ で，$\delta \in [-7.432, -7.428]$ を 0.001 刻みで変化させ，AUC が最大であった時のラフメンバシップ値によって各データポイントを色分けした結果を図 14 に示す．この時の δ は -7.430 であり，AUC は 0.7091 であった．14 から，データは主に y 軸を基準にクラスタリングが行われていることから，ユーザーが与えた評価値によってデータがグループ分けされているといえる．また，y 軸の値が 0 付近

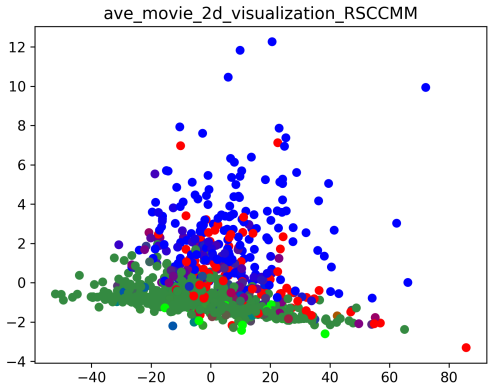

図 14: 二値化を行わない MovieLens-100k: PCA による 2 次元への視覚化

の多くのユーザーがオーバーラップして，複数のクラスターに帰属しており，y 軸が 0 より大きい値のユーザーがオーバーラップしておらず，各々唯一のクラスターに分割されている様子が確認できる．

5.5 NEEDS-SCAN/PANEL データセットにおける帰属クラスター数の観察

NEEDS-SCAN/PANEL データセットに，クラスター数 $C \in \{20, 40, 60, 80, 100\}$ で提案法 (RSCCMM-CF) を適用し，それぞれで高い AUC となったパラメータ設定での各ユーザーの帰属クラスター数をヒストグラムで視覚化したグラフを図 15，16，17，18，19 に示す．なお，結果は $\delta \in [-10.0, -5]$ を 0.5 刻みで変化させ，それぞれ 10 回試行のうち AUC が最大であった際の結果を採用した．各結果の採用されたパラメータ δ と AUC の関係を表 1 に示す．

表 1: 各クラスター数とパラメータ，AUC の関係

C	20	40	60	80	100
δ	-6.5	-8.5	-8.5	-9.5	-10.0
AUC	0.8539	0.8542	0.8553	0.8569	0.8572

図 15，16，17，18，19 から，$C = 40$ 以上での結果において，全クラスターに帰属しているユーザーが増え，ユーザーの総数に対して最も大きい割合を占めていることが確認でき，全クラスター数での結果において，唯一のクラスターにのみ帰属するユーザーも一定数存在することが確認できる．また，クラスター数を大きくしていくと，全クラスターに帰

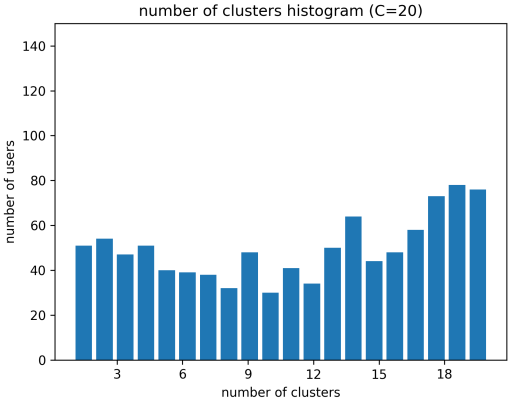

図 15: NEEDS-SCAN/PANEL: 帰属クラスター数のヒストグラム

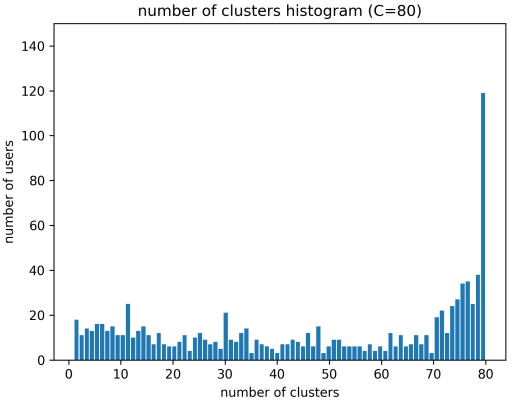

図 18: NEEDS-SCAN/PANEL: 帰属クラスター数のヒストグラム

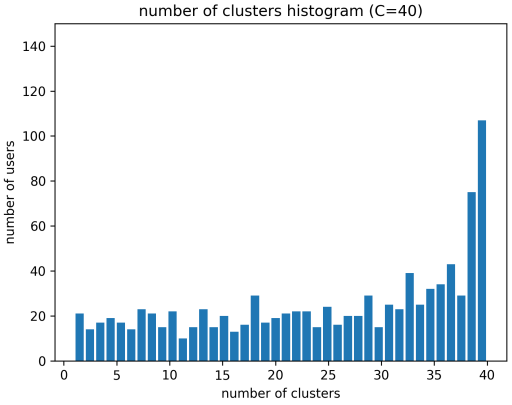

図 16: NEEDS-SCAN/PANEL: 帰属クラスター数のヒストグラム

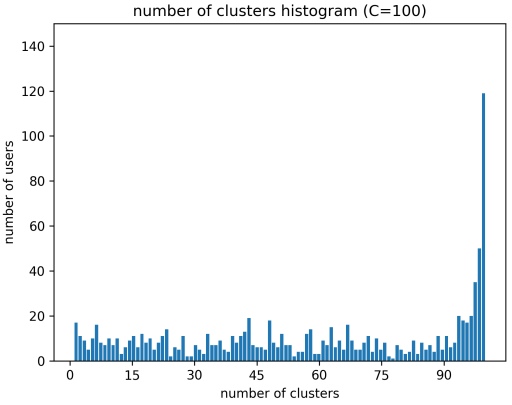

図 19: NEEDS-SCAN/PANEL: 帰属クラスター数のヒストグラム

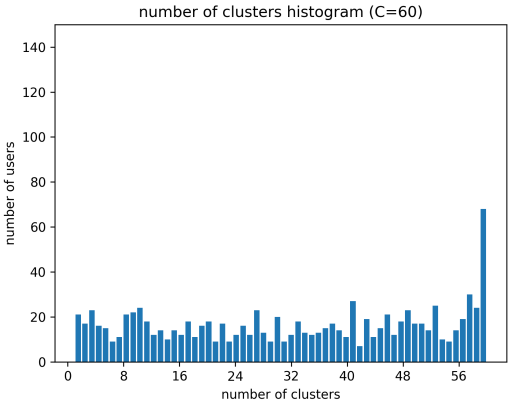

図 17: NEEDS-SCAN/PANEL: 帰属クラスター数のヒストグラム

属しているユーザーは増えていくのに対して，それ以外の一部のクラスターに帰属するユーザーの総数の変化は比較的小さいことが確認できる．

6 考察

6.1 MovieLens-100k データセットに含まれる未評価値について

MovieLens-100k データセットは全体の約 94% が未評価値という非常に疎なデータであり，未評価値の処理が実験の結果に大きく影響する．本研究ではデータの抽出と二値化を行い，未評価値を 0.5 とする処理と，二値化を行わず，未評価値を各ユーザーの平均評価値で置き換える 2 種類の処理を行った．それらのデータに提案法の RSCCMM-CF を適用した結果を比較すると，まず，$C = 1$ としてクラスタリングを行わず，各映画の評価値の平均値が全

ユーザーに対する推薦度とする設定での AUC は二値化を行った MovieLens-100k データセットの場合は 0.6940 で，二値化を行わない MovieLens-100k データセットの場合は 0.7036 となっており，二値化を行わない MovieLens-100k データセットでの AUC が約 0.0094 ほど高い結果となった．これは二値化を行うことで，本来 5 段階評価と未評価という 6 種類の状態を持つデータが，高い評価と低い評価，未評価値という 3 種類の状態を持つデータに変換されたことで情報量が減少し，各映画の評価値の平均値を全体の推薦とした際の AUC の低下につながったと考えられる．

一方で, 図 6, 8 で確認できるように, クラスター数による AUC の変化から，各 AUC の最大値は二値化を行った MovieLens-100k データセットでは $C = 4$ の時 0.7108 で，二値化を行わない MovieLens-100k データセットでは $C = 10$ の時 0.7083 であり，二値化を行った MovieLens-100k データセットでの AUC が約 0.0025 ほど高い結果となった．また，提案法を用いることで AUC が $C = 1$ の時の結果から，二値化を行った MovieLens-100k データセットでは約 0.0168 の向上が見られたのに対して，二値化を行わない MovieLend-100k データセットでは約 0.0047 しか向上していないことがわかる．これは二値化を行わない MovieLens-100k データセットでは未評価値を各ユーザーの平均評価値で補完したことによってユーザー内で似たような評価値を持つアイテムが非常に多くなり，提案法を適用するとクラスタリングの際に補完した値がユーザー固有の特徴抽出に大きく影響を与えてしまったと考えられる．つまり未評価値処理の段階でデータが既にクラスタリングされた後のような状態になってしまい，提案法の適用による AUC の向上が小さくなってしまったと考えられ，今回の提案法においてデータの前処理では未評価値に何かしらのバイアスを加えるべきではないと考えられる．

6.2 各データセットでの AUC の変化について

図 4 から，NEEDS-SCAN/PANEL データセットではクラスター数の増加に伴って HCCMM-CF を適用すると AUC が低下し，提案法を適用すると AUC が向上し続ける挙動がみられる．これはデータのアイテム数が 18 と少なく，自動車や家電などの多くの世帯が所有しているアイテムが含まれるため，ハードなクラスタリングを基にした協調フィルタリングを行うと，主要なアイテムを共通して持つユーザー同士も分割されてクラスタリングされるため，データの特徴を捉えられずに AUC が低下してしまうと考えられる．一方で，クラスタリングにおけるユーザーのオーバーラップを可能にする提案法を適用すると, 主要なアイテムを持つユーザーがオーバーラップされ，データの特徴を損なわず，AUC を向上させることができると考えられる．また，クラスタリング結果における帰属クラスター数の観察の実験結果 (図 15, 16, 17, 18, 19) からクラスター数を増加させると，すべてのクラスターの帰属するユーザーの割合が増え，その他のユーザーの割合は相対的に低くなっていくことがわかる．したがって提案法では，すべてのクラスターに帰属するユーザーには全体の所有有無の平均に近いような推薦が行われ，一部のクラスターに帰属するユーザーにはそれぞれのクラスターの特徴に応じた推薦がなされることで AUC が向上したと考えられる．

図 6，8 から，MovieLens-100k データセットではクラスター数の増加に伴って HCCMM-CF を適用すると AUC が低下し，提案法を適用すると AUC が向上し，最大値となったのち低下していく挙動がみられる．また，各パラメータ設定が NEEDS-SCAN/PANEL データセットでは $\delta \in [-10.0, -5.0]$ を 0.5 刻み，二値化を行った MovieLens-100k データセットでは $\delta \in [-8.1, -6.5]$ を 0.2 刻み，二値化を行わない MovieLens-100k データセットでは $\delta \in [-7.434, -7.428]$ を 0.001 刻みとなっており，高い AUC を得るパラメータ設定が比較的 MovieLens-100k データセットの方が限定的であることがわかる．これは直感的に MovieLens-100k データセット内にクラスター構造とオーバーラップが存在し，その構造に近づくことで AUC が向上し，過剰な分割やオーバーラップを行うと AUC が低下すると考えられる．

7 おわりに

本研究では共クラスタリング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 RSCCMM 法に基づいた協調フィルタリング手法である RSCCMM-CF を提案し，実データである NEEDS-SCAN/PANEL データセットおよび MovieLens-100k

データセットに適用し，推薦性能の変化を観察した．

実験結果から，NEEDS-SCAN/PANEL データセットにおいてパラメータを適切に設定することで提案法の RSCCMM-CF が HCCMM-CF および RCCMM-CF より高い推薦性能を持ち，クラスター数を増加させると，データを平均的な嗜好パターンを持つユーザーと特徴的な嗜好パターンを持つユーザーに二分するように機能することが確認された．

また，MovieLens-100k データセットにおいてもパラメータを適切に設定することで提案法の RSCCMM-CF が HCCMM-CF および RCCMM-CF より高い推薦性能を持つことが確認でき，データの欠損値の処理においては各ユーザーの平均値で補完したデータよりも，二値化を行って一律 0.5 で補完したデータの方が，提案法を適用した際の推薦性能の向上が大きく現れることが確認された．

以上から，ラフ集合理論の観点を導入しない HCCMM-CF や粒状化を考慮していない RCCMM-CF よりも高い推薦性能が得られたことで，ラフ集合理論の粒状性は共起関係データの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，提案法をベースとした目的関数やパラメータ設定機構の導入を試みることによって数値的な妥当性の分析を行い，さらに有効な協調フィルタリング手法を提案することも期待できる．

8 謝辞

本研究は大阪公立大学大学院情報学研究科の生方誠希准教授，本多克宏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, Proc. 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, S. Takahashi, A. Notsu, and K. Honda: Basic Consideration of Collaborative Filtering Based on Rough C-Means Clustering, Proc. of Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems, 256-261 (2020)

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

[8] H. Kim, S. Ubukata, A. Notsu, and K. Honda: Two Types of Collaborative Filtering Based on Rough Membership C-Means Clustering, Proc. of 22nd International Symposium on Advanced Intelligent Systems, 1-6 (2021)

[9] 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)

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

[11] S. Ubukata, K. Mouri, and K. Honda: Basic Consideration of Collaborative Filtering Based on Rough Co-clustering Induced by Multinomial Mixture Models, Proc. of 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), 1-6 (2022)

[12] 野岳 就拓， 生方 誠希， 野津 亮， 本多 克宏: ラフ集合理論に基づく粒状性を考慮した共クラスタリングに関する一検討, インテリジェント・システム・シンポジウム講演論文集，354-359 (2021)