

# 修士学位論文

---

## 題 目

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 Granularity

---

主査	宇野 裕之	教授
副査	本多 克宏	教授
副査	内海 ゆづ子	講師
副査	生方 誠希	准教授

令和 5 年 ( 2023 年 ) 度修了  
(No. BGA22050) 毛利 憲竜

---

大阪公立大学大学院情報学研究科

基幹情報学専攻 知能情報学分野

# 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 Granularity

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 などの動画配信サービス等のコンテンツ推薦システムで広く活用されている。CF においては、アイテムベース CF [1] やベイジアン CF [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 CF 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がある。

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 (ユーザー) 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、CF 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、各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 (RMCMM) 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 CF が提案されている [6, 7, 8]。

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を抽出する共クラスタリングが注目されている。CF で扱うデータはユーザー  $\times$  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 法の特例である Hard CCMM (HCCMM) 法をベースとして、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した CF (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 HCCMM 法に基づく CF (HCCMM-CF) および RCCMM 法に基づく CF

(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 CF 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 6 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), 共クラスタリング手法 (HCCMM 法, RCCMM 法, RSCCMM 法), および RCCMM-CF について概説し, 第 3 章では提案法である RSCCMM-CF を説明する。第 4 章では数値実験の設定と結果, 第 5 章で考察を示す。最後に, 第 6 章で本論文のまとめを述べる。

## 2 準備

$C$ -Means 型クラスタリングでは,  $n$  個の対象からなる全体集合  $U = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$  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  $\mathbf{x}_i = (x_{i1}, \dots, x_{ij}, \dots, x_{im})^\top$  で表されるとし, 各クラスター  $c$  はクラスター中心  $\mathbf{b}_c = (b_{c1}, \dots, b_{cj}, \dots, b_{cm})^\top$  を持つとする。

### 2.1 HCM 法

HCM 法は代表的な非階層的クラスタリング手法であり, 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  $\mathbf{b}_c$  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。

$$d_{ci} = \|\mathbf{x}_i - \mathbf{b}_c\|, \quad (1)$$

$$d_i^{\min} = \min_{1 \leq l \leq C} d_{li}, \quad (2)$$

$$u_{ci} = \begin{cases} 1 & (d_{ci} \leq d_i^{\min}) \\ 0 & (\text{otherwise}). \end{cases} \quad (3)$$

**Step 4** クラスター中心  $\mathbf{b}_c$  を計算する。

$$\mathbf{b}_c = \frac{\sum_{i=1}^n u_{ci} \mathbf{x}_i}{\sum_{i=1}^n u_{ci}}. \quad (4)$$

**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。したがって, 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 法にラフ集合理論の観点を導入し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, 対象の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, GRM 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 GRM with Membership Normalization (GRM-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  $\alpha$  ( $\alpha \geq 1$ ),  $\beta$  ( $\beta \geq 0$ ) を設定する。

**Step 2**  $C$  個の初期クラスター中心  $\mathbf{b}_c$  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  $\bar{u}_{ci}$  と正規化メンバシップ  $\tilde{u}_{ci}$  を順に以下の式で求める。

$$\bar{u}_{ci} = \begin{cases} 1 & (d_{ci} \leq \alpha d_i^{\min} + \beta) \\ 0 & (\text{otherwise}), \end{cases} \quad (5)$$

$$\tilde{u}_{ci} = \frac{\bar{u}_{ci}}{\sum_{l=1}^C \bar{u}_{li}}. \quad (6)$$

**Step 4** クラスター中心  $\mathbf{b}_c$  を計算する。

$$\mathbf{b}_c = \frac{\sum_{i=1}^n \tilde{u}_{ci} \mathbf{x}_i}{\sum_{i=1}^n \tilde{u}_{ci}}. \quad (7)$$

**Step 5**  $\bar{u}_{ci}$  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：

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

RSCM 法は、RCM 法と同様に明確な目的関数を持たず、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  $b_c$  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  $b_c$  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  $\mu_{ci}^R$  と上近似に対するメンバシップ  $\bar{u}_{ci}$  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。

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

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

**Step 5** クラスター中心  $b_c$  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

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, 2 に共クラスタリングの例を示す。

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

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

	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 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。

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

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

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

クラスター  $c$  と対象  $i$  の類似度  $s_{ci}$  を次式で定める。

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

ここで、一般に  $r_{ij}$  は非負として、 $s_{ci} \leq 0$  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、それぞれ総和が 1 となるように正規化する。

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

**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。

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

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

**Step 4** 項目メンバシップ  $w_{cj}$  を更新する。

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

**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、対象の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  $\alpha$  ( $\alpha \geq 1$ )、 $\beta$  ( $\beta \leq 0$ ) 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  $\bar{u}_{ci}$  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。

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

**Step 4** 項目メンバシップ  $w_{cj}$  を更新する。

$$w_{cj} = \frac{\sum_{i=1}^n \bar{u}_{ci} r_{ij}}{\sum_{l=1}^m \sum_{i=1}^n \bar{u}_{ci} r_{il}}. \quad (20)$$

**Step 5**  $\bar{u}_{ci}$  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、 $\alpha$  が大きな値を取るほど、また  $\beta$  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。

## 2.4.3 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。 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  $\mu_{ci}^R$  と上近似に対するメンバシップ  $\bar{u}_{ci}$ , 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.

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

**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.

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

## 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:

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

対象  $i$  と対象  $t$  の類似度  $S_{it}$  を下記のように定義する:

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

これは, 負の交差エントロピーとみなせる. 交差エントロピーは 2 つの確率分布が離れている度合いを示すため,  $S_{it}$  は類似度とみなせる. 類似度  $S_{it}$  に基づき, 二項関係を以下のように設定する:

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

ここで,  $\delta \leq 0$  はラフさを調節するパラメータであり,  $\delta$  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  $\delta$  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 (NSP) 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), ワークプロ (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), オープン (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 (ML) 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 ML データセットと二値化を行わない ML 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 ML 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

表 1: 混同行列

実際	予測	
	Positive	Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

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  $r_{ij} = 1$  に, 3 以下であれば  $r_{ij} = 0$  に置き換え, 未評価値は  $r_{ij} = 0.5$  とする変換を行った.

二値化を行わない ML 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) は偽陽性率 (false positive rate, FPR) に対する真陽性率 (true positive rate, TPR) であり, AUC (Area Under the ROC Curve) は ROC 曲線の下側の面積である. ここで, 偽陽性率は実際陽性でないものが陽性であると判断された割合, 真陽性率は実際陽性であるものが陽性であると判断された割合のことである. 混同行列を表 1 に示す. 真陽性率および偽陽性率は (26) 式のように計算される.

$$TPR = \frac{TP}{TP + FN}, \quad FPR = \frac{FP}{FP + TN} \quad (26)$$

推薦が完全にランダムに行われた場合 AUC は 0.5 に近い値となり, AUC が 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 NSP データセット

HCCMM-CF, RCCMM-CF および提案法である RSCCMM-CF を NSP 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  $\delta$  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

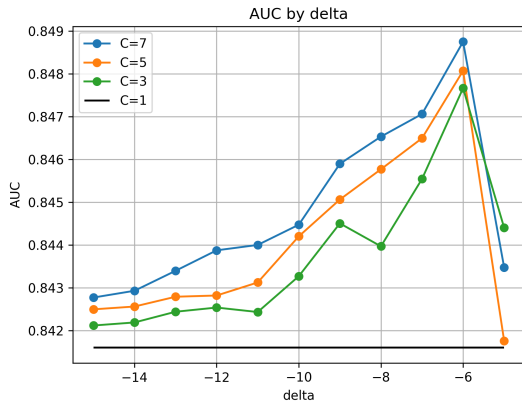


図 3: NSP: 各  $C$  における  $\delta$  による AUC の変化 (RSCM-CF)

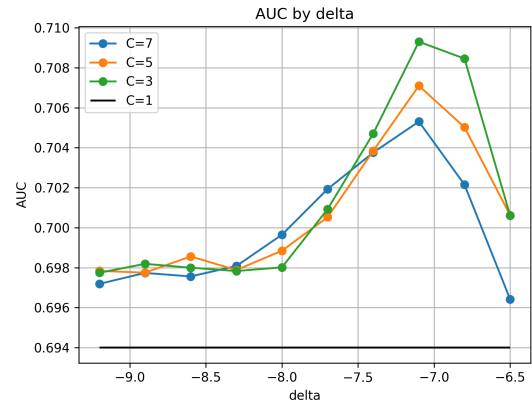


図 5: 二値化を行った ML: 各  $C$  における  $\delta$  による AUC の変化 (RSCM-CF)

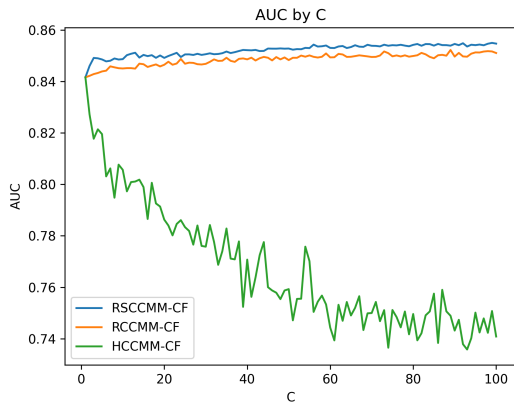


図 4: NSP:  $C$  による AUC の変化 (RSCM-CF, RCCM-CF, HCCM-CF)

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。  $\delta$  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界に対する推薦度となり、  $\delta$  の値に依存せず、 AUC は 0.8416 である。  $C = \{3, 5, 7\}$  の場合、  $\delta$  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。これにより、CF タスクにおけるクラスタリングの有効性および粒状化による不確実性の取り扱いの有効性が確認できる。

次に HCCM-CF, RCCM-CF および提案法である RSCM-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、HCCM-CF は異なる初期値による 5 回試行の平均値であり、RSCM-CF では  $\delta \in [-10.0, -5.0]$  を 0.5 刻み、RCCM-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、RCCM-CF および RSCM-CF が HCCM-CF より高い AUC を持つことがわかる。したがって、クラスターのオーバーラップを考慮することが CF タスクにおいて有効であることがわかる。さらに、RSCM-CF が RCCM-CF より高い AUC を持つことから、粒状性に基づく不確実性の取り扱いが有効であることがわかる。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 HCCM-CF の AUC は低下する反面、RCCM-CF および RSCM-CF の場合は安定した AUC を持つことも確認できる。

#### 4.3.2 二値化を行った ML データセット

HCCM-CF, RCCM-CF および RSCM-CF を二値化を行った ML 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  $\delta$  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、  $\delta$  の値に依存せず、AUC は 0.6940 である。図 5 から、



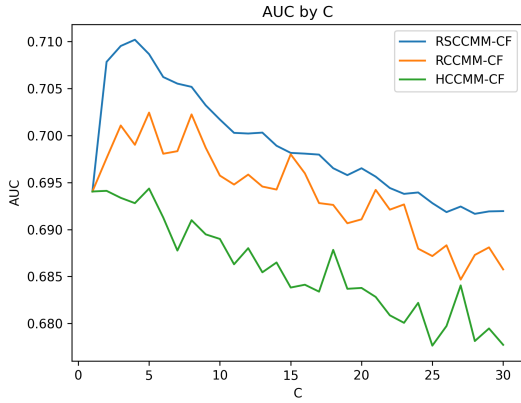


図 6: 二値化を行った ML:  $C$  による AUC の変化 (RSCMM-CF, RCCMM-CF, HCCMM-CF)

$C = \{3, 5, 7\}$  の場合,  $\delta$  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. これにより, CF タスクにおけるクラスタリングの有効性および粒状化による不確実性の取り扱いの有効性が確認できる.

次に HCCMM-CF, RCCMM-CF および RSCMM-CF について, クラスタ数  $C$  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, RSCMM-CF では  $\delta \in [-8.1, -6.5]$  を 0.2 刻み, RCCMM-CF では  $\beta$  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, クラスタ数に関わらず, RCCMM-CF および RSCMM-CF が HCCMM-CF より高い AUC を持つことがわかる. したがって, クラスタのオーバーラップを考慮することが CF タスクにおいて有効であることがわかる. さらに, RSCMM-CF が RCCMM-CF より高い AUC を持つことから, 粒状性に基づく不確実性の取り扱いが有効であることがわかる. また, クラスタ数  $C$  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  $C = 5$ , RSCMM-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.

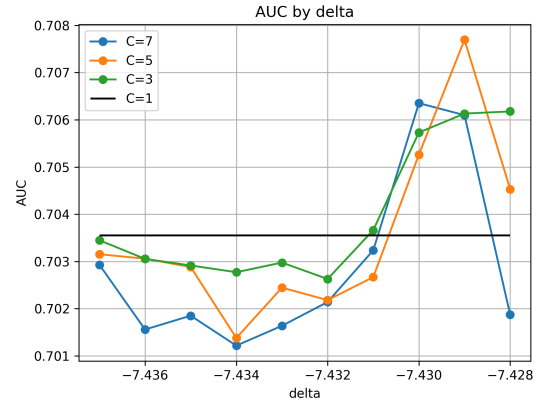


図 7: 二値化を行わない ML: 各  $C$  における  $\delta$  による AUC の変化 (RSCMM-CF)

#### 4.3.3 二値化を行わない ML データセット

HCCMM-CF, RCCMM-CF および提案法である RSCMM-CF を二値化を行わない ML 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  $\delta$  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,  $\delta$  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,  $\delta$  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. これにより, CF タスクにおけるクラスタリングの有効性および粒状化による不確実性の取り扱いの有効性が確認できる.

次に HCCMM-CF および RSCMM-CF について, クラスタ数  $C$  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, RSCMM-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

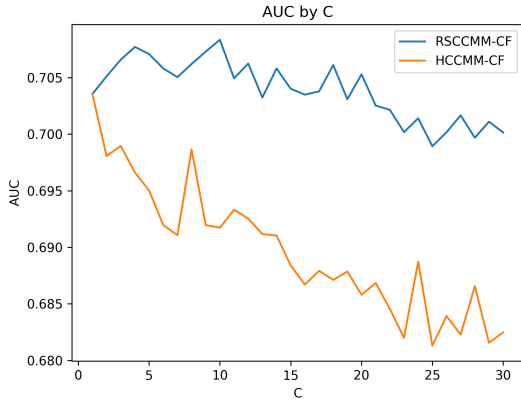


図 8: 二値化を行わない ML:  $C$  による AUC の変化 (RSCCMM-CF, HCCMM-CF)

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、RSCCMM-CF が HCCMM-CF より高い AUC を持つことが確認できる。これにより、粒状性に基づく不確実性の取り扱いが有効であることがわかる。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。

#### 4.4 各データセットの視覚化と提案法の適用

NSP, 二値化を行った ML, 二値化を行わない ML の各データセットの評価値行列に RSCCMM 法を適用し、主成分分析 (Principal Component Analysis, 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。

##### 4.4.1 NSP データの視覚化

NSP 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。PCA によって得られた主成分について検証を行うと、第一主成分の固有ベクトルは、中・小電気冷蔵庫を除いて負の値をとっており、多くの製品を所有すると第一主成分得点は減少することが確認できた。特に、パソコンやピアノなどの比較的値段が高い製品の固有ベクトルの絶対値が大きく、これらを所有していると第一主成分得点は小さくなり、所有している製品の数が少ないかつ中・小電気冷蔵庫を所有

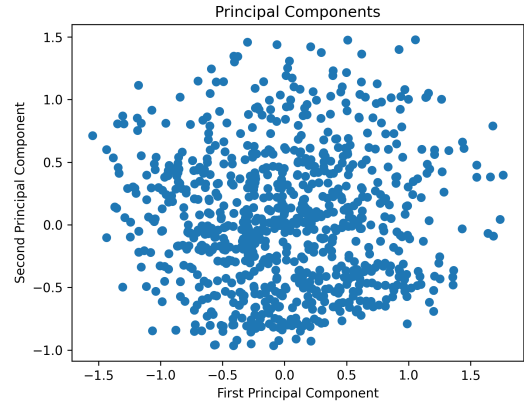


図 9: NSP: PCA による 2 次元への視覚化

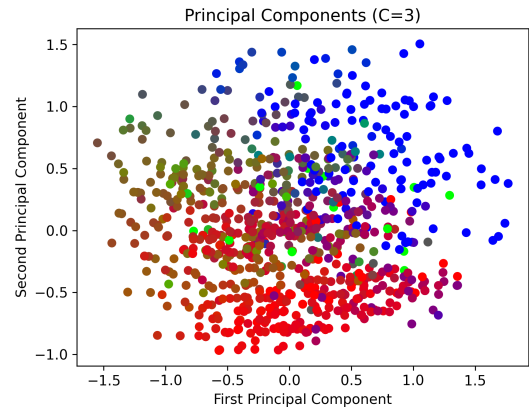


図 10: NSP: RSCCMM 法によるクラスタリング結果の視覚化 ( $C = 3$ )

していると、第一主成分得点は大きくなることがわかる。このことから、様々な製品の購入に積極的な世帯がグラフの左側、そうでなく最低限の製品を購入する世帯が右側に位置していると考えられる。また、第二主成分の固有ベクトルで絶対値が大きい製品を確認すると、負の値で大型電気冷蔵庫や電気乾燥機、正の値でワープロや中・小電気冷蔵庫、VD などであると確認できた。したがって、製品の特性から、比較的世帯人数の多い大規模世帯がグラフの下側、小規模世帯がグラフの上側に位置していると考えられる。

そして、 $C = 3$  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ー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  $\delta$  は  $-6.0$  であり、AUC は 0.8515 であった。図 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

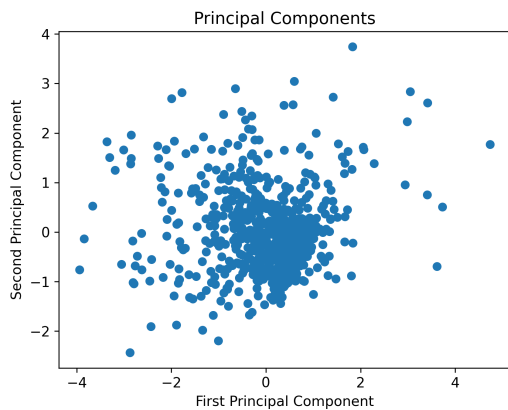


図 11: 二値化を行った ML: PCA による 2 次元への視覚化

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、その他のデータは比較的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。

#### 4.4.2 二値化を行った ML データの視覚化

二値化を行った ML 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。PCA によって得られた主成分について検証を行い、第一主成分得点が最も高い 10 ユーザーの平均評価値は 0.4000、低い 10 ユーザーの平均評価値は 0.6184 と大きく差があることが確認できた。このことから、映画に低い評価をつける傾向のあるユーザーがグラフの右側、高い評価をつける傾向のあるユーザーがグラフの左側に位置していると考えられる。そして、第二主成分が最も高い 10 ユーザーの未評価値の数の平均値は 326.8、低い 10 ユーザーの未評価値の数の平均値は 444.4 と、こちらも大きく差があることが確認できた。したがって、評価した映画数が比較的多いユーザーがグラフの上側、少ないユーザーがグラフの下側に位置していると考えられる。

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  $\delta$  は  $-6.9$  であり、AUC は 0.7125 であった。図 12 から、クラスターが右上、右下、左側に分布しており、原点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

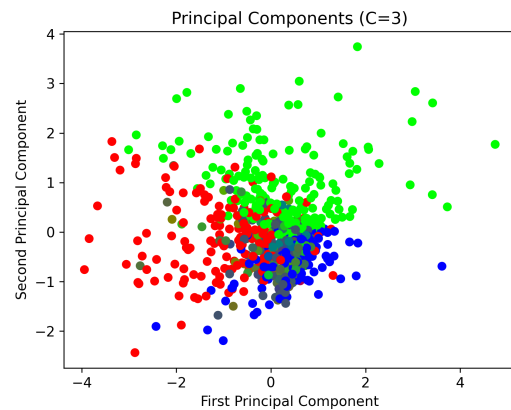


図 12: 二値化を行った ML: RSCMM 法によるクラスタリング結果の視覚化 ( $C = 3$ )

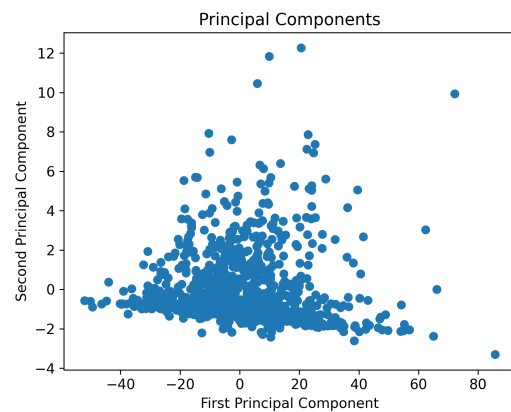


図 13: 二値化を行わない ML: PCA による 2 次元への視覚化

しており、原点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。

#### 4.4.3 二値化を行わない ML データの視覚化

二値化を行わない ML 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 によって得られた主成分について検証を行い、第一主成分得点が最も高い 10 ユーザーの平均評価値は 2.0603、低い 10 ユーザーの平均評価値は 4.6718 と大きく差があることが確認できた。このことから、映画に低い評価をつける傾向のあるユーザーがグラフの右側、高い評価をつける傾向のあるユーザーがグラフの左側に位置しているといえる。そして、第二主成分が最も高い 10

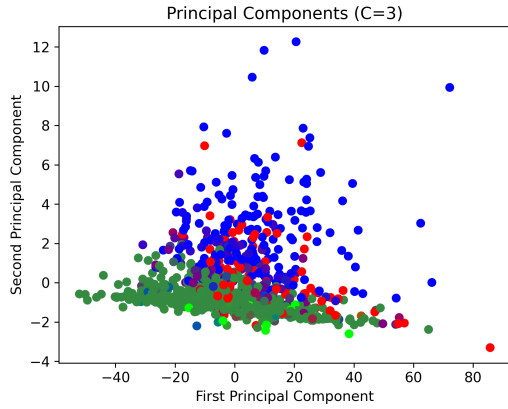


図 14: 二値化を行わない ML: RSCCMM 法によるクラスタリング結果の視覚化 ( $C = 3$ )

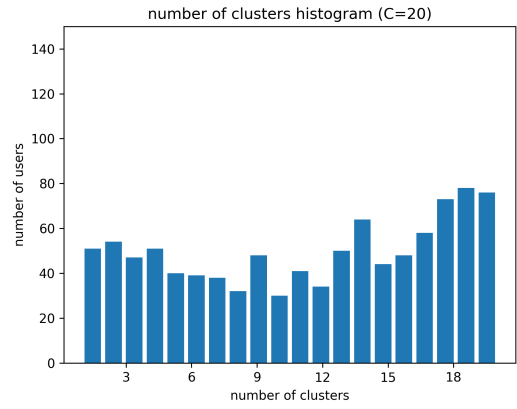


図 15: NSP: 帰属クラスター数のヒストグラム ( $C = 20$ )

表 2: 各クラスター数とパラメータ, AUC の関係

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

ユーザーの未評価値の数の平均値は 1253.5, 低い 10 ユーザーの未評価値の数の平均値は 1578.1 と, こちらも大きく差があることが確認できた. したがって, 評価した映画数が比較的多いユーザーがグラフの上側, 少ないユーザーがグラフの下側に位置しているといえる.

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  $\delta$  は -7.430 であり, AUC は 0.7091 であった. 図 14 から, データの分割は  $x$  軸では明確ではないが, -10 付近でユーザー分割の傾向がみられる.  $y$  軸では, 0 付近かそれより小さい値のユーザーの多く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比較的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.

#### 4.5 NSP データセットにおける帰属クラスター数の観察

NSP 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

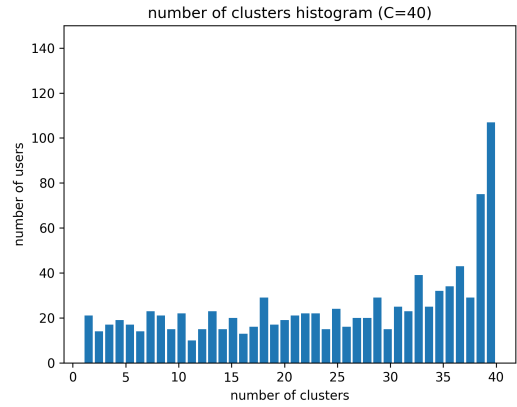


図 16: NSP: 帰属クラスター数のヒストグラム ( $C = 40$ )

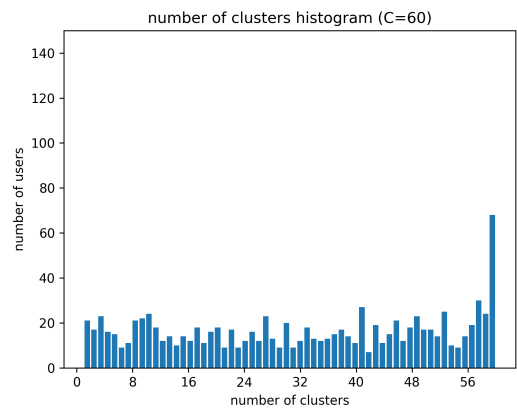


図 17: NSP: 帰属クラスター数のヒストグラム ( $C = 60$ )

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

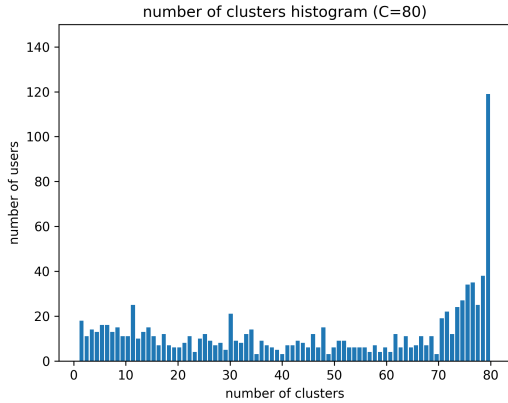


図 18: NSP: 帰属クラスター数のヒストグラム  
( $C = 80$ )

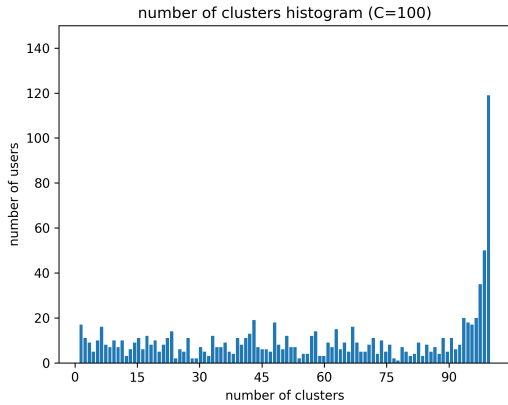


図 19: NSP: 帰属クラスター数のヒストグラム  
( $C = 100$ )

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。各結果において採用されたパラメータ  $\delta$  と AUC の関係を表 2 に示す。

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加し、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る。一方で、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、クラスター数を増加させると、全クラスターに帰属するユーザーが増加するのに対して、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。

## 5 考察

### 5.1 NSP データセットでの AUC の変化について

図 4 から、NSP データセットでは HCCMM-CF を適用するとクラスター数の増加に伴って AUC が低下し、提案法や RCCMM-CF を適用すると AUC が向上し続ける挙動がみられる。また、図 10 から、クラスターのオーバーラップが比較的グラフの左側に位置する世帯で多く発生していることが確認できる。

つまり、PCA の検証を踏まえると、多くの種類の製品を所有する世帯がオーバーラップしやすいということがわかる。これらはデータの製品数が 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製品が含まれるため、ハードなクラスタリングを基にした CF を行うと、主要な製品を共通して持つ世帯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や RCCMM-CF を適用すると、多くの種類の製品を持つ世帯にオーバーラップが起こりやすいという結果から、主要な製品を持つ世帯にオーバーラップが起こりやすくなり、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。

さらに、提案法ではユーザー同士の関連性から粒状化を行うことで、クラスターとの類似度でオーバーラップを判断する RCCMM-CF に比べてよりユーザー同士の関連性をクラスター分割に反映させることができ、RCCMM-CF よりも高い AUC を実現できたと考えられる。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に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。

このことから、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の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。



## 5.2 ML データセットでの AUC の変化について

図 6, 8 から, ML データセットでは HCCMM-CF を適用するとクラスター数の増加に伴って 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 NSP 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 ML 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 ML 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 ML データセットの方が限定的であることがわかる. これらから, ML データセット内にオーバーラップをもつクラスター構造が存在し, 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 が低下したと考えられる.

## 5.3 ML データセットの未評価値処理について

ML 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,  $C = 1$  としてクラスターリングを行わず, 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 ML データセットの場合は 0.6940 で, 二値化を行わない ML 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 ML 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,  $C = 1$  として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 ML データセットでは  $C = 4$  の時 0.7108

で, 二値化を行わない ML データセットでは  $C = 10$  の時 0.7083 であり, 最大値を比較すると二値化を行った ML 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による AUC の変化量の最大値を見ると,  $C = 1$  の時の結果から, 二値化を行った ML 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 ML 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.

これらの原因としては下記の 2 点が考えられる. 1 点目は二値化を行った ML 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  $m = 583$  の映画を抽出して実験を行っており, ユーザーごとの情報量は二値化を行わない ML データセットより多くなっていった点であると考えられる. したがって, 二値化を行ったことでそれぞれの評価値の差についての情報量は減っていたとしても, 抽出によってクラスター分割の指標となるユーザーごとの情報量は大きくなり, AUC の増加が二値化を行わない ML データセットよりも大きくなったと考えられる. このことは, 図 12 にみられるクラスター分割で, ユーザーの平均評価値の影響を強く受ける第一主成分 ( $x$  軸方向) でのクラスター分割が比較的是っきりと表れていることから確認できる.

2 点目は二値化を行わない ML 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映画が非常に多くなり, 二値化を行って未評価値を 0.5 で補完した場合に比べて, 未評価値とそうでない値の差が小さくなってしまい, ユーザー固有の特徴を捉えることが難しくなってしまった点だと考えられる. 提案法は各ユーザーの特徴ベクトルを確率分布として捉えるため, ユーザー内の各アイテムの共起度の違いが各ユーザーの特徴抽出に非常に重要であり, 未評価値が多いデータにおいて各ユーザーの平均値で補完したことが影響を強く与えたと考えられる. 図 14 にみられるクラスター分割においても, ユーザーの平均評価値の影響を強く受ける第一主成分 ( $x$  軸方向) に基づいた分割が, 図 12 にみられるクラスター分割に比べて不明瞭であり,  $y$  軸方向に基づいた分割が強くみられ, クラスター分割において, ユーザーの平均評価値よりも未評価値の数がクラスター分割に強く影響していることが確認できる.

したがって、今回の提案法においてデータの事前処理では、ユーザー内のアイテム同士の共起度の違いが明確になるような処理が求められると考えられる。

## 6 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、粒状性を考慮したラフ共クラスタリング手法である RSCCMM 法に基づく協調フィルタリング手法として 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である RSCCMM-CF が従来法の 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である RSCCMM-CF が従来法の HCCMM-CF や RCCMM-CF より高い推薦性能を持つことが確認できた。また、データの欠損値の前処理においては各ユーザーの平均値で補完した場合よりも、二値化を行って一律 0.5 で補完した場合の方が、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、提案法が、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、ラフ集合理論に基づく粒状性の考慮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、今回の提案法にあった欠損値の前処理やクラスタリング結果の分析を発展させ、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。今後の課題としては、適切なパラメータの決定基準の導入などが挙げられる。

## 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。また、有用な助言を多数いただいた同大学大学院現代システム科学研究科の野津亮教授に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, #OS03-2, 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, #F-2-C-1, 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)