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

人間情報システム研究グループ　毛利 憲竜

1. はじめに

電子商取引サイトや動画配信サービス等に見られるコンテンツ推薦システムは協調フィルタリング (Collaborative Filtering, CF) により実現され，推薦性能を高めることでユーザビリティや企業利益の向上が見込める．クラスタリングベースの CF では，嗜好の類似したユーザーからなるクラスターを抽出し，各ユーザーに対し，それぞれが帰属するクラスター内で嗜好度の高いコンテンツを推薦する．

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と考えられる．したがって，ラフ集合理論 [1] 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．また，CF におけるデータはユーザー × 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．

共クラスタリング手法として，対象の分割に関してハードな Hard Co-Clustering induced by Multinomial Mixture models (HCCMM) 法やラフ集合理論の観点を導入した Rough CCMM (RCCMM) 法 [2] があり，それに基づく CF (RCCMM-CF)[3] が提案されている．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．そこで，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 [4] が提案されている．

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 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．

2. RSCCMM 法に基づく協調フィルタリング

2.1. 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 を二項関係 $R \subseteq U \times U$ によって粒状化し，対象のクラスターへの帰属を粒ごとに判定することで，帰属の不確実性を取り扱い，クラスターのオーバーラップを実現する．R の設定の仕方によって多様な粒状化がなされ，多様な分類が可能となる．

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 c に対するメンバシップを w_{cj}，対象数を n，項目数を m として，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 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．

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

Step 3 クラスター c と対象 i の類似度を s_{ci} とし，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．

$$s_{ci} = \sum_{j=1}^{m} r_{ij} \log w_{cj}, \tag{2}$$

$$s_i^{\max} = \max_{1 \le c \le C} s_{ci}, \tag{3}$$

$$u_{ci} = \begin{cases} 1 & (s_{ci} \ge s_i^{\max}), \\ 0 & (\text{otherwise}). \end{cases} \tag{4}$$

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値 $\tilde{u}_{ci}$ を順に以下の式で計算する．

$$\mu_{ci}^R = \frac{\sum_{t=1}^{n} R_{it} u_{ct}}{\sum_{t=1}^{n} R_{it}} \tag{5}$$

$$\overline{u}_{ci} = \begin{cases} 1 & \left(\mu_{ci}^R > 0\right), \\ 0 & (\text{otherwise}). \end{cases} \tag{6}$$

$$\tilde{u}_{ci} = \frac{\overline{u}_{ci}}{\sum_{l=1}^{C} \overline{u}_{li}}. \tag{7}$$

Step 5 項目メンバシップ w_{cj} を以下の式で更新する．

$$w_{cj} = \frac{\sum_{i=1}^{n} \tilde{u}_{ci} r_{ij}}{\sum_{l=1}^{m} \sum_{i=1}^{n} \tilde{u}_{ci} r_{il}}. \tag{8}$$

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

2.2. 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．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，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:

$$\tilde{r}_{ij} = \frac{r_{ij}}{\sum_{l=1}^{m} r_{il}}. \tag{9}$$

共起関係データに適した類似度を考えるため，((2) 式) を参考に，対象 i と対象 t の類似度 S_{it} を下記のように定義する:

$$S_{it} = \sum_{j=1}^{m} \tilde{r}_{tj} \log \tilde{r}_{ij}. \tag{10}$$

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:

$$R_{it} = \begin{cases} 1 & (S_{it} \ge \delta), \\ 0 & (\text{otherwise}). \end{cases} \tag{11}$$

ここで，$\delta \le 0$ 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．

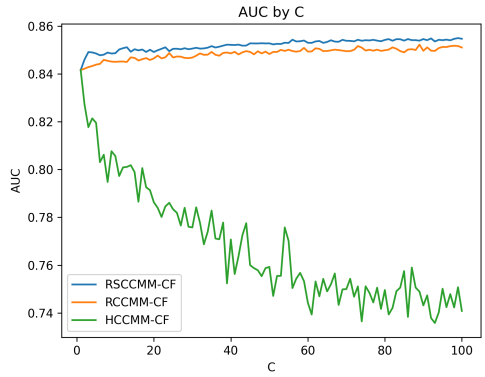

図 1: NEEDS-SCAN/PANEL データ：各手法における C による AUC の変化

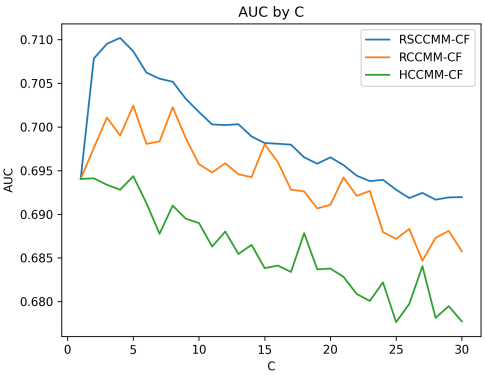

図 2: 二値化を行った MovieLens-100k データ：各手法における C による AUC の変化

2.3. RSCCMM-CF

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 (11) 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 $\hat{r}_{ij}$ を以下の式で計算する．

$$\hat{r}_{ij} = \sum_{c=1}^{C} \tilde{u}_{ci} w_{cj}. \tag{12}$$

Step 3 閾値 $\eta \in [\min\{\hat{r}_{ij}\}, \max\{\hat{r}_{ij}\}]$ を設定し，閾値以上の推薦度を持つアイテムを以下の式のように推薦する．

$$\check{r}_{ij} = \begin{cases} 1 & (\hat{r}_{ij} \geq \eta)\,, \\ 0 & (\text{otherwise})\,. \end{cases} \tag{13}$$

3. 数値実験

3.1. 実験概要

3 種類の実データ (NEEDS-SCAN/PANEL, 二値化を行った MovieLens-100k，二値化を行わない MovieLens-100k) に対して提案法 (RSCCMM-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パラメータ δ による推薦性能の変化を検証した．評価指標として ROC-AUC を用いた．また，従来法 (HCCMM-CF, RCCMM-CF) との比較，提案法を適用した際の発生したクラスターの分析を行った．

3.2. 実験結果

NEEDS-SCAN/PANEL データにおいて初期クラスター数 C を 1 から 100 の範囲で変化させたときの AUC の変化を図 1 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 1 から，クラスター数に関わらず，提案法の RSCCMM-CF が RCCMM-CF および HCCMM-CF より高い AUC を持つことがわかる．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 の AUC は低下する反面，RCCMM-CF および RSCCMM-CF の場合は安定した AUC を示すことも確認できる．

次に二値化を行った MovieLens-100k データにおいて，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 2 に示す．各 AUC は，HCCMM-CF で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 2 から，クラスター数に関わらず，提案法の RSCCMM-CF が RCCMM-CF および 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．

4. 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，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．

参考文献

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

[2] 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)

[3] 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)

[4] 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)