凸包を用いた協調フィルタリング

Collaborative Filtering with Convex Hull

経営システム工学専攻
菊間　優希

1.　序論

近年，インターネットなどの情報通信技術の発達は目覚ましく，日々膨大なデータが生まれており，これらはビッグデータと呼ばれている．このようなビッグデータを活用するうえで，ユーザーにとって有用と思われる対象，情報，または商品などを選び出しそれをユーザーの目的に合わせた形で提示する，情報推薦システムが注目されている．情報推薦システムの代表的な手法に協調フィルタリングがあり，ユーザーの行動履歴と類似する履歴を持つユーザーの評価を利用して推薦する．

推薦システムの礎となった GroupLens は協調フィルタリングの内でも，利用者間メモリベース法と呼ばれる．GroupLens の方法は，二段階あり，第一段階に嗜好が似ているといえる，類似度の高いユーザーを求める．類似度とは，嗜好パターンがどれくらい似ているかを定量化したものである．第二段階で，推薦対象者が知らないアイテムについて，第一段階で求めた類似度に基づいて推薦対象者がどれくらいそのアイテムを好むかを予測する．第一段階の類似度に着目するだけでも，様々な手法，改良が研究されている．

そこで本研究では，新たな類似度の指標として Dula, Bougnol [2] により提案された，凸包を用いた類似ユーザーの特定手法 ゲージ LP を基に，新たな手法を提示する． ゲージ LP では扱うデータによって実行不可能となるため，実行可能範囲へ射影することで，データの性質に依らない類似ユーザーの特定を行う．数値実験では，提案手法によって求められた類似ユーザーと実際の評価値との予測精度により評価する．

2.　凸包を用いた類似ユーザーの特定手法

Dula, Bougnol [2] により提案された，凸包を用いた類似ユーザーの特定手法についてまとめる．

2.1.　協調フィルタリングの流れ

以下の用語を用いて，メモリベース法である協調フィルタリングの評価予測の流れを説明する．

まず，アイテム・ユーザーの分類を示す．

- ターゲットユーザー (target user) ：特定のアイテムの評価が不明で，類似ユーザーの評価を使用して予測されるユーザー
- ターゲットアイテム (target item) ：ターゲットユーザーの評価が不明なアイテム
- ピアユーザー (peer users) ：ターゲットユーザーによって評価されたアイテムの評価と，ターゲットアイテムの評価を持つユーザー
- 類似ユーザー (similar users)：アイテムのスコアが，ターゲットユーザーのスコアに近いユーザー

下記の図 2.8 のようにピアユーザー J を $j = 1, \cdots, n$ とおき，新規顧客のアイテムに対する評価のうち，ターゲットアイテムを除いた m 種のアイテム評価を $b = (b_1, \cdots, b_m) \in \mathbb{R}^m$ とおく．$a_j \in \mathbb{R}^m$ はターゲットユーザーが既に評価を付けているアイテムのピアユーザー j の評価値ベクトルである．α_j はピアユーザーのターゲットアイテムの値を示す．β はターゲットユーザーの予測したいターゲットアイテムの値を示す．

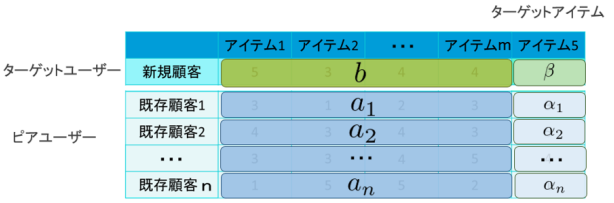


図 2.1: ユーザー・アイテム行列値の対応

各 b に対し，β を予測する流れを以下にまとめる．

1. 類似ユーザー集合 $S_b \subset \{1, \cdots, n\}$ を見つける
2. $\alpha_j, \alpha_j \in S_b$ から β を予測する

2.2.　凸包を用いた類似ユーザーの特定手法

Dula, Bougnol [2] ではピアユーザの中からターゲットユーザの類似ユーザを特定するために，次の線形計画問題（LP）を解くことを提案している．:

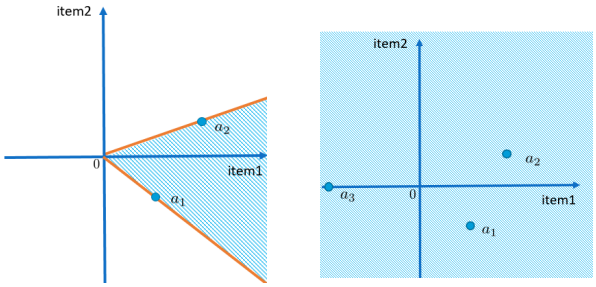
$$
\text{(GLP)} \qquad f(b) := \left| \begin{array}{ll} \min\limits_{\lambda} & \sum\limits_{j=1}^{n} \lambda_j \\ \text{s.t.} & \sum\limits_{j=1}^{n} a_j \lambda_j = b, \\ & \lambda_j \geq 0, \quad j = 1, \ldots, n. \end{array} \right.
$$

LP (3.1) を [2] に倣ってゲージ LP（Gauge LP，以下 GLP）と呼ぶ．GLP を幾何的に解釈するために，次の概念を導入する．

定義 1 (非負包) ベクトル $a_1, ..., a_n \in \mathbb{R}^m$ に対し，以下で与えられる集合を $a_1, ..., a_n$ の非負包 (non-negative hull) と呼ぶ．

$$NN(\{a_1, ..., a_n\}) := \Big\{x : x = \textstyle\sum_{j=1}^{n} \lambda_j a_j, \\ \lambda_j \geq 0, j = 1, ..., n\Big\}$$

たとえば $a_1 = (1, -1), a_2 = (2, 1), a_3 = (-2, 0)$ に対し非負包 $NN(\{a_1, a_2\})$ は図 2.2 のようになり，非負包 $NN(\{a_1, a_2, a_3\})$ は $\mathbb{R}^2$ に一致し，2 次元平面全体となる．定義から $NN(\{a_1, ..., a_n\})$ は錐であり，非空な集合に対する非負包は常に原点を含む，つまり $A \neq \emptyset$ であれば $0 \in NN(A)$．また原点が $\{a_1, ..., a_n\}$ の凸包の内部に含まれる場合，$NN(\{a_1, ..., a_n\})$ は全空間 $\mathbb{R}^m$ になる．



全空間 $\mathbb{R}^2$ にならない例　　全空間の例

図 2.2: 非負包の例

この非負包の概念を用いると，GLP ピアユーザの非負結合としてターゲットユーザが表現できるとして，そのときの非負重みの和を最小にする問題である．これの幾何的解釈を深めてみる．

図 2.2 の右は 3 人のピアユーザの凸包が原点がを含む場合を例示している．このとき目的関数は凸包のスケーリングサイズに対応しており，最適値はターゲットユーザがスケーリングした凸包に含まれる最小のサイズを表す．実際，ターゲットユーザが凸包の (a) 内部 (1 より小さい)，(b) 境界上 (1 に等しい)，(c) 外部 (1 より大きい) にいるかどうかが対応している．

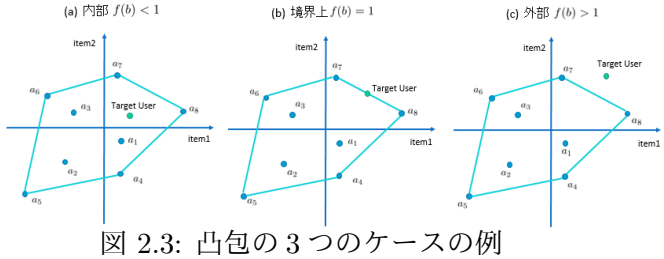


図 2.3: 凸包の 3 つのケースの例

最適解を $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$ とおくと $\lambda_j^* > 0$ なるピアユーザ j をターゲットユーザの類似ユーザーとする．さらに類似ユーザーであると判断された場合の λ_j^* の値が大きいほどターゲットユーザーに近いユーザーであると判断できる．

一方で，ピアユーザー $a_1, \cdots, a_n$ の凸包が原点を含まない場合には非負包の全空間の真部分集合になるが，ターゲットユーザーが含まれていないので GLP は実行不可能である．この場合，GLP は類似ユーザの特定が不可能となる．Dula, Bougnol [2] ではこのような場合を明示していないため，いくつかの対処法を提示する．

3. 凸包を用いた類似ユーザーの特定における改訂手法

GLP が実行不可能であった場合にピアユーザーの非負包上で最も近い点を見つけ，その点をターゲットユーザーの代理として GLP を実行する対処法を提示する．

3.1. ℓ_2 ノルムによる射影

GLP が実行不可能であった場合, ピアユーザーの非負包上で最も近い点を見つける際に，ℓ_2 ノルムを用いる方法を考える．以下の凸 2 次計画問題（QP）を解くことを考える：

$$\mathrm{Proj}_{\ell_2} \quad \left| \begin{array}{ll} (f^{**})^2 = \min\limits_{\lambda,\varepsilon} & \|\varepsilon\|_2^2 \\ \text{s.t.} & \sum\limits_{j=1}^{n} a_j \lambda_j + \varepsilon = b, \\ & \lambda_j \geq 0, \quad j = 1, \ldots, n. \end{array} \right.$$

Proj_{ℓ_2} は制約式に残差ベクトル $\varepsilon \in \mathbb{R}^m$ を導入することで必ず最適解が存在する．ここで，最適値の正の平方根をとったものを f^{**}，最適解を $(\lambda, \varepsilon) = (\lambda^{**}, \varepsilon^{**})$ とおく．f^{**} はターゲットユーザーの非負包からの距離を表す値と見なせる．$b - \varepsilon^{**}$ は，ピアユーザーの非負包上の点となり，これをターゲットユーザーの代理として，類似ユーザーを特定することを考える．その上で以下の修正した GLP を解くことを考える．

$$\mathrm{MGLP}_{\ell_2} \quad \left| \begin{array}{ll} \min\limits_{\lambda} & \sum\limits_{j=1}^{n} \lambda_j \\ \text{s.t.} & \sum\limits_{j=1}^{n} a_j \lambda_j = b - \varepsilon^{**}, \\ & \lambda_j \geq 0, \quad j = 1, \ldots, n. \end{array} \right.$$

以下の図 3.1 に非負包の外部にいるターゲットユーザーを ℓ_2 ノルムで射影させる概念図を示す．

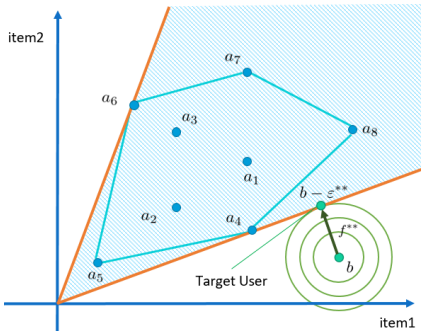


図 3.1: ℓ_2 ノルムによる射影の概念図

3.2. ℓ_1 ノルムによる射影

Proj_{ℓ_2} は QP であったが，ℓ_2 ノルム（の平方）を ℓ_1 ノルムに置き換えることで，LP に基づく同様の問題へと帰着可能である．実際，ℓ_1 ノルムを用いれば問題は次の LP に帰着される：

$$\mathrm{Proj}_{\ell_1} \left| \begin{array}{ll} f^* := \min\limits_{\lambda,\varepsilon^+,\varepsilon^-} & \sum\limits_{i=1}^{m}(\varepsilon_i^+ + \varepsilon_i^-) \\ \text{s.t.} & \sum\limits_{j=1}^{n} a_j\lambda_j + \varepsilon^+ - \varepsilon^- = b, \\ & \lambda_j \geq 0, \qquad j = 1,\ldots,n, \\ & \varepsilon^+, \varepsilon^- \geq 0, \ i = 1,\ldots,m. \end{array} \right.$$

以下の図 3.2 に非負包の外部にいるターゲットユーザーを ℓ_1 ノルムで射影させる概念図を示す．

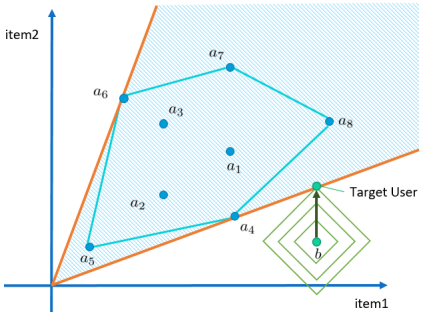


図 3.2: ℓ_1 ノルムによる射影の概念図

この最適値を f^* とすると，代理ベクトルは $b-(\varepsilon^{*+}-\varepsilon^{*-})$ と表される．この代理ベクトルに対する修正 GLP を以下のように与える．

$$\left| \begin{array}{ll} \min\limits_{\lambda,\varepsilon^+,\varepsilon^-} & \sum\limits_{j=1}^{n}\lambda_j \\ \text{s.t.} & \sum\limits_{j=1}^{n} a_j\lambda_j + \varepsilon^+ - \varepsilon^- = b, \\ & \sum\limits_{i=1}^{m}(\varepsilon_i^+ + \varepsilon_i^-) = f^*, \\ & \lambda_j \geq 0, \qquad j = 1,\ldots,n, \\ & \varepsilon^+, \varepsilon^- \geq 0, \quad i = 1,\ldots,m. \end{array} \right.$$

3.3. ℓ_∞ ノルムによる射影

ℓ_∞ ノルムに代えて，ℓ_∞ ノルムを用いても近接点問題は LP に帰着できる．実際，以下の LP によってもターゲットユーザーの代理をピアユーザーの非負包上にみつけることができる．

$$\mathrm{Proj}(\ell_\infty) \left| \begin{array}{ll} f^{***} := \min\limits_{s,\lambda,\varepsilon^+,\varepsilon^-} & s \\ \text{s.t.} & \varepsilon_i^+ + \varepsilon_i^- \leq s, i = 1,\cdots,m, \\ & \sum\limits_{j=1}^{n} a_j\lambda_j + \varepsilon^+ - \varepsilon^- = b, \\ & \lambda_j \geq 0, \qquad j = 1,\ldots,n. \end{array} \right.$$

以下の図に非負包外にいるターゲットユーザーを ℓ_∞ ノルムで射影させる概念図を示す．

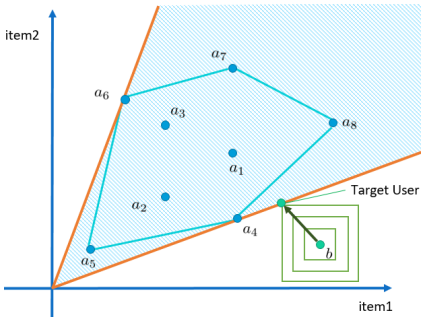


図 3.3: ℓ_∞ ノルムによる射影の概念図

この最適値を f^{***} とすると，ターゲットユーザーの代理となるベクトルは $b-(\varepsilon^{+***}-\varepsilon^{-***})$ と表される．この代理ベクトルに対する GLP は以下のように与えられる．

$$\left| \begin{array}{ll} \min\limits_{\lambda} & \sum\limits_{j=1}^{m}\lambda_j \\ \text{s.t.} & \sum\limits_{j=1}^{n} a_j\lambda_j = b - (\varepsilon^{+***} - \varepsilon^{-***}), \\ & \lambda_j \geq 0, \qquad j = 1,\ldots,n. \end{array} \right.$$

4. 数値実験

数値実験を行い，提案手法によって求められた類似ユーザーと実際の評価値との予測精度（平均絶対誤差）を示す．

4.1. 実験に使用した実データとデータのスパース性について

数値実験には Jester[1]，MovieLens[3] を用いた．実験データに未評価値がある場合は非負包が定義ができず，QP や LP が意味をなさないため，未評価部分を補完する，実際に用いたアルゴリズム [4] を以下に示す．

$$\mathrm{MC} \left| \begin{array}{ll} \underset{U}{\text{minimize}} & \mathrm{rank}(U) \\ \text{subject to} & U_{ij} = A_{ij} \quad \forall (i,j) \in \rho \end{array} \right.$$

ただし $\rho \subset \{(i,j) : 1 \leq i \leq m, 1 \leq j \leq n\}$. ユーザー・アイテム行列を補完した．補完した上で各データセットのピアユーザー数を変化させ，学習

データと検証データに分けたものを 24 種類作成した．比較対象となる類似度を測る指標は，ユークリッド距離，コサイン距離，相関距離を用いた．

4.2. 計算時間の比較

図 4.1 は計算時間の比較を表したものである．

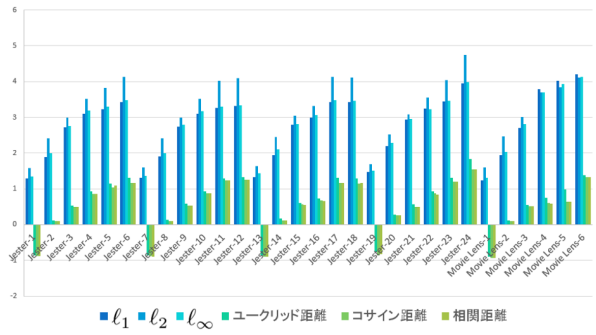


図 4.1: 計算時間の比較

計算時間は比較手法が圧倒的に早い一方で，ユーザー一人当たりの計算時間は長くても 1 分程度で推薦を行える．したがって，問題の構造を生かしたアルゴリズムを構築することで，計算効率にかかわる提案手法の実用性は高められると考えられる．

4.3. 予測精度の比較

図 4.2，4.3 はそれぞれ予測精度について比較したものである．

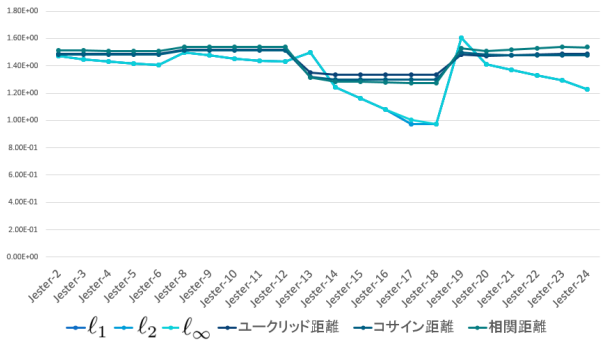


図 4.2: Jester の予測精度の比較

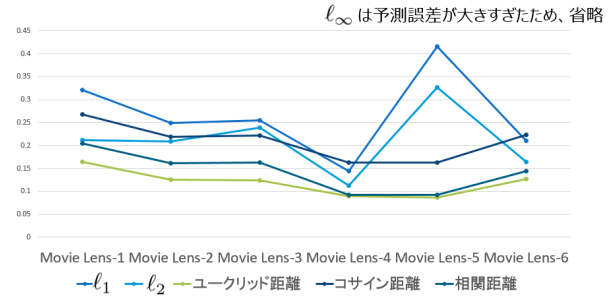


図 4.3: Movie Lens の予測精度の比較

Jester のデータセットではピアユーザー数が多いほど比較手法と比べて提案手法の予測精度が高い結果となった．しかし，Movie Lens のデータセットでは比較手法の方が精度が高い結果となった．アイテム数がピアユーザー数よりはるかに大きい Jester に比べ，MovieLens は比較的アイテム数とピアユーザー数の差が大きくなく，類似ユーザーを求める際のアイテム数に対応する次元数の高さによる精度の低下だと考えれらる．

5. 結論

本論文では，凸法を用いた類似ユーザーの特定手法 [2] に対して，ノルムによる射影を用いた，代理ユーザーを求める方法を提案した．実行不可能範囲にターゲットユーザーが位置する場合に，ℓ_1，ℓ_2，ℓ_∞ ノルムを用いて実行可能範囲への近接点を求めることで類似ユーザーの特定が可能となった．

今回試した実データでは，計算時間は他手法に比べかかってしまうが，他手法では求められなかった，類似ユーザーの重み（重要度）を用いることで，予測精度を高めることができることを示せた．

今後の課題としては，アイテム数とユーザー数が均衡しているようなデータに対する予測精度の向上, データの欠損値に対する補完手法の検討，計算時間の短縮等があげられる．

参考文献

[1] D. Goldberg, D. Nichols, Brian M. Oki, and D. Terry. “Using collaborative filtering to weave an information tapestry,” Commun. ACM, Vol. 35, No. 12, pp. 61–70, December 1992.

[2] H. Dula and L. Bougnol. “The Recommender Problem with Convex Hulls,” ISMP Bordeaux, スライド, July 5, 2018.

[3] Movielens. University of Minnesota. `http://movielens.org/`, 最終閲覧 2019 年 2 月 26 日.

[4] Z. Wen, W. Yin, Y. Zhang. “Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm,” Mathematical Programming Computation, Vol. 4, Issue 4, pp. 333–361, 2012