

Collaborative Metric Learning に基づく推薦リストの意外性向上手法に関する研究

1X18C098-7 米田安希子
指導教員 後藤正幸

1. 研究背景・目的

近年、ユーザの属性や利用履歴からユーザ個人の嗜好を把握し、嗜好に合致したアイテムを提示する推薦システムは重要な Web マーケティング技術の 1 つとなっている。従来の推薦システムは、ユーザがアイテムに対して直接的な評価値を与える明示的評価を対象としていたが、近年では評価値が直接的に与えられないクリックや閲覧などの暗黙的評価の活用が期待されている。しかし、ユーザの暗黙的評価が観測されたアイテム（以下、観測アイテム）はユーザの好みに合致すると考えられるが、暗黙的評価が観測されていないアイテム（以下、未観測アイテム）は単にユーザに認知されていないアイテムであるのか、ユーザの好みと合致しないために未観測であるのかを区別することが困難である。そこで、暗黙的評価を対象とした推薦モデルとして Collaborative Metric Learning（以下、CML）[1] が知られている。

CML では、異なるアイテムに対するユーザの相対的な嗜好の差を捉えることに焦点を置き、ユーザは未観測アイテムよりも観測アイテムを相対的に好むと仮定し、未観測アイテムよりもユーザの近くに埋め込むように学習する。この学習により、ユーザとアイテムの関係性に加えて、ユーザ同士、アイテム同士の類似性などを埋め込み空間で表現することができる。さらに、学習した CML における埋め込み表現を用いることで、ユーザとの距離が近いアイテムを推薦効果が高いものとして推薦することが可能になる。

しかし、CML では観測数が多いアイテム（以下、メジャーアイテム）ほど学習時に重要視される傾向があり、メジャーアイテムに偏重した推薦が行われてしまうという問題がある。このようなユーザの嗜好に関係のないメジャーアイテムに偏重した推薦は、ユーザにとっての意外性の低下につながるため、効果が低いと考えられる。すなわち、観測数に関係なく、ユーザの嗜好に沿ったアイテムを推薦することが、ユーザにとっては好ましい。そこで本研究では、ユーザの嗜好を捉え、観測数が少ないアイテム（以下、マイナーアイテム）も含むような意外性のある推薦を可能とする手法を提案する。具体的には、メジャーアイテムであるほど損失が小さくなるように観測数に応じた重み付けを行うことで、メジャーアイテムに偏重しない推薦が可能となる。最後に、実際の映画評価データに提案手法を適用し、ユーザの嗜好を考慮した意外性のある推薦における提案手法の有用性を示す。

2. Collaborative Metric Learning

2.1. CML の概要

CML は、ユーザとアイテムを同一空間上に埋め込むモデルであり、ユーザは未観測アイテムよりも観測アイテムを相対的に好むと仮定して、未観測アイテムよりもユーザの近くに埋め込むように学習する。これにより、CML は暗黙的評価におけるユーザの嗜好、ユーザやアイテム間の類似性を表

現することができる。

2.2. CML の定式化

CML では、ユーザと各アイテムのユークリッド距離の近さが、各アイテムの相対的な嗜好の順位と対応するように学習が行われる。いま、ユーザ i とアイテム j の埋め込み表現をそれぞれ $\mathbf{u}_i, \mathbf{v}_j$ 、ユーザ i と観測アイテム k のペア (i, k) の集合を \mathcal{S} 、ユーザ i と未観測アイテム ℓ のペア (i, ℓ) の集合を $\bar{\mathcal{S}}$ 、 \mathbf{u}_i と \mathbf{v}_j のユークリッド距離を $d(i, j) = \|\mathbf{u}_i - \mathbf{v}_j\|$ として、CML の損失関数を式 (1) に示す。

$$L_m(d) = \sum_{(i,k) \in \mathcal{S}} \sum_{(i,\ell) \in \bar{\mathcal{S}}} w_{ik} [m + d(i, k)^2 - d(i, \ell)^2]_+ \quad (1)$$

ただし、 $[x]_+ = \max(0, x)$ とする。式 (1) では、ユーザ i と観測アイテム k の距離が未観測アイテム ℓ よりも、予め定めたマージン距離 m 以上遠い場合のみ損失とみなしている。また、 w_{ik} はユーザ i と観測アイテム k のペア (i, k) に対する重みであり、式 (2) に示す Weighted Approximate Rank Pairwise 損失 [2] を用いる。

$$w_{ik} = \log(\text{rank}_d(i, k) + 1) \quad (2)$$

式 (2) は、全観測アイテムにおける観測アイテム k とユーザ i の距離の近さの順位 $\text{rank}_d(i, k)$ に基づいた重みを表し、順位の低い観測アイテムの重みが大きくなる。これにより、多くの未観測アイテムよりもユーザとの距離が相対的に遠いであろう順位の低い観測アイテムとの関係性も考慮した学習が可能となる。CML の学習では、ミニバッチ確率的勾配降下法により式 (1) の損失を最小化することでユーザおよびアイテムの埋め込み表現 $\mathbf{u}_i, \mathbf{v}_j$ を推定する。

また、CML で学習した埋め込み空間から推薦を行う際には、各ユーザに対して埋め込み表現のユークリッド距離が近いアイテムを順に M 個選択し、推薦リストを作成する。

3. 提案手法

3.1. 概要

一般的な評価履歴データでは、全アイテムの大半をマイナーアイテムが占めている。しかし CML を用いた推薦では、ユーザ i と観測アイテム k のペア (i, k) の集合 \mathcal{S} にメジャーアイテムが多く含まれており、学習時にこれらのアイテムの埋め込み表現をユーザに近づけようとするため、メジャーアイテムに偏重した推薦がされてしまう。そこで本研究では、ユーザの嗜好に沿った意外性のあるアイテムを含めた推薦を行うことを目的とし、メジャーアイテムであるほど損失への影響が小さくなるように、観測数の差異を考慮した重みを損失関数に導入する新たな手法を提案する。

3.2. 提案手法の定式化

CML での埋め込み空間における損失 $L_m(d)$ に、観測アイテム k に対する重み α_k を考慮した新たな損失関数 $L_{m2}(d)$ を式 (3) に定義する。

$$L_{m2}(d) = \sum_{(i,k) \in S} \alpha_k \sum_{(i,\ell) \in S} w_{ik} [m + d(i,k)^2 - d(i,\ell)^2]_+ \quad (3)$$

ただし、観測アイテム k に対する重み α_k は観測数 X_k を用いて、式 (4) で定義される。ここで、 a, b は事前に設定するハイパーパラメータである。

$$\alpha_k = aX_k^{-b} \quad (4)$$

この重み付けにより、メジャーアイテムになるほど損失が小さくなるため、式 (3) の最小化では偏重した学習が解消される。そのため、メジャーアイテムばかり推薦されることなく、ユーザの嗜好に合致したマイナーアイテムも含まれた意外性のある推薦が可能になると期待される。

4. 実データ分析

提案手法の有効性を示すため、従来手法としての CML と推薦精度の比較を行う。

4.1. 分析条件

本実験では、5 段階の映画評価データである MovieLens を分析のために利用するが、本研究の対象は暗黙的評価であるため、評価値 4, 5 のデータを暗黙的評価における観測データ、それ以外を未観測データとして扱う。観測数 $N = 575, 129$ 、ユーザ数 $U = 6, 038$ 、映画数 $I = 3, 381$ であり、観測データを 3 : 1 : 1 の割合で、学習データ、検証データ、テストデータに分割した。また、重み α_k のハイパーパラメータについては、 $a = 1$ とし、 b を 0.1 から 1 まで変化させた。評価については、各ユーザへの推薦アイテム数 M を 50 と設定した。また、テストデータに含まれる各ユーザの観測データのうち、推薦リストに含まれるデータの割合で定義される再現率 (Recall) を活用し、各ユーザの再現率の平均を評価指標として算出した。マイナー映画に絞った推薦を行う際は、マイナー映画のみを対象として算出した再現率 ($Recall_{minor}$) を用いる。評価は、学習結果 10 回の再現率の平均により行う。

4.2. マイナー映画に絞った推薦の結果と考察

提案手法がマイナー映画に対するユーザの嗜好を捉えているかを確認するために、全学習データにより学習し、マイナー映画のみを推薦する場合の提案手法の性能を評価する。ここでは、マイナー映画を推薦する比較手法として、従来の CML を用いた以下の 2 つの方法を用いる。

従来手法 (1) 全映画のデータを学習し、推薦時にマイナー映画のみをリスト化する方法

従来手法 (2) 学習時にマイナー映画のデータのみを学習し、推薦する方法

マイナー映画と定義する観測数の閾値を 10 件未満から 500 件未満へと変化させた結果を図 1 に示す。

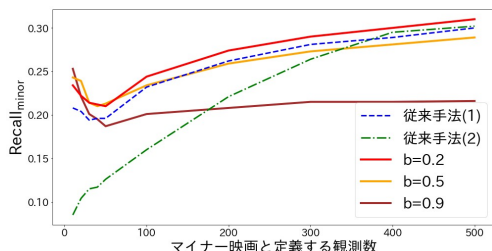


図 1: マイナー映画における再現率

図 1 より、提案手法 ($b = 0.2$) は各従来手法より再現率が向上している。したがって、提案手法 ($b = 0.2$) はマイナー映画に対する嗜好をより捉えているといえる。また、観測数の少ない映画に絞るほど、 b の値が大きい提案手法の再現率が高くなる。そのため、重み α_k の損失に対する影響を大きくするほど、よりマイナーな映画を学習時に重視するようになると考えられる。

4.3. 全映画に対する推薦の結果と考察

全観測データを用いて学習し、全映画から推薦リストを作成した結果を表 1 に示す。ここで、*,** はそれぞれ $p \leq 0.01$, $p \leq 0.001$ で統計的に有意であることを示す。

表 1: 全映画における再現率

	従来手法	$b = 0.1$	$b = 0.2$	$b = 0.3$
再現率	0.360	0.365*	0.370**	0.360

表 1 より、従来手法の再現率は 0.360、提案手法 ($b = 0.2$) の再現率は 0.370 であり、全映画から推薦する場合でも、提案手法がよりユーザの嗜好に沿った推薦を行っているといえる。また、推薦リストに含まれる映画のうち従来手法と提案手法 ($b = 0.2$) で 39.1% の映画が異なっており、重み付けにより埋め込み空間上の位置関係が変化しているといえる。次に、従来手法と提案手法 ($b = 0.2$) において、各映画の観測数を横軸、推薦リスト中で各ユーザによって実際に観測された映画の全ユーザでの累積数を縦軸として図 2 に示す。

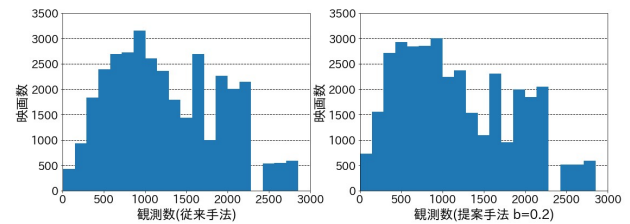


図 2: 実際に観測された映画の観測数

図 2 より、提案手法ではユーザの嗜好に沿ったマイナー映画をより多く含むことが確認された。したがって、提案手法ではよりユーザの嗜好を捉えた学習を行い、観測数の少ない映画も含むような意外性のある推薦が可能であるといえる。

5. まとめと今後の課題

本研究では、CML がメジャーアイテムに偏重して学習してしまうという問題に着目し、メジャーアイテムに偏重しない意外性のある推薦を目的として、観測数に応じた損失を導入した CML の拡張手法を提案した。そして、実際の映画評価データを用いた分析により、提案手法がよりユーザの嗜好に従い、かつ意外性のある推薦を行うことを示した。今後の課題として、アイテムのカテゴリなどの補助情報を考慮したモデルの構築などが考えられる。

参考文献

- [1] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge J. Belongie, Deborah Estrin, "Collaborative Metric Learning," *Proceedings of the 26th International Conference on World Wide Web*, pp.193–201, 2017.
- [2] Nicolas Usunier, David Buffoni, Patrick Gallinari, "Ranking with ordered weighted pairwise classification," *Proceedings of the 26th Annual International Conference on Machine Learning*, pp.1057–1064, 2009.