日本オペレーションズ・リサーチ学会
2022年 春季研究発表会

ノイズに頑健な協調距離計量学習

東京工業大学 *松井諒生 MATSUI Ryo
エムシーデジタル株式会社 柳沼傑 YAGINUMA Suguru
SMN 株式会社 内藤 剛人 NAITO Taketo
01405430 東京工業大学 中田和秀 NAKATA Kazuhide

1. はじめに

ユーザーの属性情報や web 上の行動履歴などをもとに，ユーザーそれぞれの好みに合った web コンテンツを自動で推薦する「推薦システム」の研究が盛んに行われている．特に，現代的な推薦システムでは主に 2 つの特徴が求められている．

1 つ目の要求は，implicit feedback と呼ばれるデータの利用である．implicit feedback とは web サイトの閲覧履歴など，ユーザーの能動的な行動を必要としないインタラクションのログデータを表す．このようなデータはデータ収集コストが小さいという利点がある一方で，データに含まれるノイズは比較的大きいと言える．そのため，implicit feedback の性質を考慮したノイズに頑健な推薦システムの構築を目指す必要がある．

2 つ目の要求は埋め込みベクトルの利用である．埋め込みベクトルとは，観測した (ユーザー—アイテム) 間の関係性データをもとに，ユーザーおよびアイテムをそれぞれ 1 つのベクトルとして表現したものである．そして，ユーザーおよびアイテムの潜在的な嗜好や類似性をそのベクトルの類似度によって予測する．近似近傍探索などの手法を用いればユーザーベクトルに類似したアイテムベクトルを，ユーザー体験の毀損が起きない時間内で発見することができる．さらに，ベクトルの類似度はユーザーの潜在的な嗜好性に合ったアイテムを捉えるだけでなく，(ユーザー–ユーザー) 間および (アイテム–アイテム) 間の類似度も表現できる．これにより，顧客セグメントを捉えるマーケティングへの応用や，SNS における友達の推薦，EC サイトにおける関連商品・代替商品の推薦など，幅広い応用が可能となる．

本発表では，以上の 2 つの要求を満たす推薦システムの構築を目指した研究 [1] について発表する．具体的には，2 点目をより正確に実行するために研究されてきた協調距離計量学習 [2] (Collaborative Metric Learning, CML) と呼ばれるアルゴリズムを拡張する．

2. 関連研究

CML は埋め込みベクトルを用いた最も標準的なアルゴリズムである行列分解が，(ユーザー–ユーザー) 間および (アイテム–アイテム) 間の細かな関係性を表現できていないという指摘を受けて開発された．CML は，ユーザー集合 $\mathcal{U}$ およびアイテム集合 $\mathcal{I}$ に対して，インタラクションを観測したユーザーアイテムペア $(u,i) \in \mathcal{S} \subset \mathcal{U} \times \mathcal{I}$ に，ユーザー u とインタラクションしていないネガティブアイテム $j \in \mathcal{U}_u \subset \mathcal{I}$ を加えたトリプレット (u,i,j) を構成単位とする目的関数

$$\mathcal{L}(\Theta) = \sum_{(u,i)\in\mathcal{S}} \sum_{j\in\mathcal{D}_u} \left[1 + d(\mathbf{v}_u, \mathbf{v}_i)^2 - d(\mathbf{v}_u, \mathbf{v}_j)^2\right]_+$$

を持つ．ただし，Θ は学習する全ての埋め込みベクトル集合 $\Theta = \{\mathbf{v}_c\}_{c\in\mathcal{U}\cup\mathcal{I}}$ を表し，$\mathcal{S}$ はインタラクションを観測したユーザーアイテムペア集合，$\mathcal{D}_u$ はユーザー u とインタラクションを観測していないアイテム集合，d はユークリッド距離，$[\cdot]_+$ は，$[x]_+ = \max\{0, x\}$ を表す．そして，全ベクトルのノルムが 1 以下という制約のもと，$\mathcal{L}(\Theta)$ を最小にする埋め込みベクトル集合 Θ を求める．その際，一様ランダムなトリプレットの抽出と勾配降下法によるパラメータ更新を繰り返すことによって実用的な局所最適解を得ることができる．距離関数は定義から三角不等式の関係性を満たすため，(ユーザー–アイテム) 間の関係性データのみから学習しても (ユーザー–ユーザー) 間および (アイテム–アイテム) 間の関係性を捉えることができると考えられている．

しかし，単純な CML は implicit feedback のノイズに対処していない．Tran らは CML の学習において部分的にノイズラベル問題へ取り組んだ [3]．ネガティブアイテム j の抽出において 2 段階のサ

ンプリング方法を提案し，ユーザーが自身の嗜好性に合ったアイテムを見逃しているという事象を指すネガティブノイズへの対策を行った．

3. 提案手法

しかしながら，ここまでの研究には依然として2つの課題がある．まず1つ目がCMLにおいて重み付けのサンプリングが十分でない可能性があるという点である．CMLが採用している距離関数の三角不等式の性質上，1度でもノイズのあるデータがサンプリングされると，ノイズの影響が全体的に広がるため学習に大きな被害を与える可能性がある．2つ目がポジティブアイテム i についてのノイズを考慮していない点である．ユーザーがあるアイテムにクリックしたものの，実際にはそのアイテムに興味がなかったという事象を指すポジティブノイズも無視できないと考えられる．

よって，この2つの問題を解消するため，本研究ではCMLの学習の前にデータの信頼度を推定し，ポジティブ・ネガティブの両ノイズに対処した重み付けまたはクリーニングに活用する方法を提案する．

3.1. 提案手法1: 双重み付けサンプリング

1つ目の提案手法である双重み付けサンプリングはトリプレット (u, i, j) の抽出において，ユーザー u のアイテム i, j に対するインタラクション $Y_{u,i} \in \{-1, +1\}$ の潜在確率 $P(Y_{u,i} = \pm 1)$ を用いて重み付けする．具体的には (u, i) の抽出確率を $P(Y_{u,i} = +1)$ で重み付け、ユーザー u に対する j の抽出確率を $P(Y_{u,j} = -1)$ で重み付ける．この $P(Y_{u,i} = \pm 1)$ は，一般的な推薦システムのアルゴリズムによって事前に推定するが，1節で示した要求を満たす必要はなく，より条件が緩いモデルを使用することができる．

3.2. 提案手法2: クリーニング

2つ目の提案手法であるクリーニングはトリプレット (u, i, j) の抽出において，$P(Y_{u,i} = +1)$ が小さい上位 p%の (u, i) の抽出確率を0，$P(Y_{u,j} = -1)$ が小さい上位 q%の (u, j) の抽出確率を0とする．そして，その他のユーザー・アイテムについては一様にサンプリングする．同様に，ここで用いる $P(Y_{u,i} = \pm 1)$ は，一般的な推薦システムのアルゴリズムによって求める．割合 p%,q%は任意に設定できるが，本研究ではどちらも5%とした．

4. 実データによる検証

本実験では，単純に一様サンプリングしたCML(*uniform*)，1段階のネガティブサンプリングをしたCML(*1stage*)，2段階のネガティブサンプリングをしたCML(*2stage*)の3つをベースラインとする．そして，提案手法の双重み付けサンプリング (*weight*) およびクリーニング (*clean*) をこれらと比較する．

表1はノイズがある環境下の推薦性能が検証可能なことで知られるYahoo! R3データセットを用いて計算した各手法の評価指標を表している．この結果の通り，双重み付けサンプリングおよびクリーニングは，全ての指標において *uniform* などのベースラインから大幅に精度が向上している．また，2つの提案手法を比較すると，クリーニングが安定的であることがわかる．

表 1: Yahoo! R3 データセットにおける推薦性能

	nDCG@3	MAP@3	Recall@3
uniform	0.494	0.550	0.523
1stage	0.450	0.493	0.502
2stage	0.416	0.459	0.464
weight	0.541	0.596	0.571
clean	**0.559**	**0.612**	**0.590**

参考文献

[1] Ryo Matsui, Suguru Yaginuma, Taketo Naito, and Kazuhide Nakata. Confident Collaborative Metric Learning. In *Proceedings of the IEEE International Conference on Data Mining Workshops (ICDMW)*, pp. 246–253, 2021.

[2] Cheng-Kang Hsieh, Longqi Yang, Yin Cui, Tsung-Yi Lin, Serge Belongie, and Deborah Estrin. Collaborative metric learning. In *Proceedings of the 26th International Conference on World Wide Web*, pp. 193–201, 2017.

[3] Viet-Anh Tran, Romain Hennequin, Jimena Royo-Letelier, and Manuel Moussallam. Improving collaborative metric learning with efficient negative sampling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1201–1204, 2019.