1N-01

Soft Actor-Critic 強化学習を用いた系列推薦フレームワーク

洪 惠珍 †　木村 優介 ‡　波多野 賢治 §
† 同志社大学文化情報学部　‡ 同志社大学大学院文化情報学研究科　§ 同志社大学文化情報学部

1 はじめに

推薦システムにおいて推薦性能を向上させるためには，ユーザの行動パターンを把握する必要がある．近年は，ユーザの行動を系列データとして捉え，ユーザがあるサービスを利用すれば何らかのログデータが追加される．系列推薦システムは，このログデータを学習に利用可能な強化学習の特徴を利用して，ユーザの嗜好変化を把握できるという理由から，ユーザの行動変化を学習できる仕組みとして注目されている [1].

強化学習は，価値関数を利用して選択した行動が報酬にどの程度影響を与えたかを学習し評価する Value-Based アルゴリズムと，報酬が多くなる行動を学習する Policy-Based アルゴリズムに分けられるが，近年では双方の利点を利用する Actor-Critic (AC) アルゴリズムがよく利用されている．AC アルゴリズムには，Actor が最適な行動を選び，Critic はその行動の価値を評価することで，相互に作用しながら改善する機能がある [2]．そのため，ユーザの行動変化を連続的な行動空間内で表現できる AC アルゴリズムは，系列推薦システムにおいてうまく動作すると言われている [3].

しかし，推薦システムで AC アルゴリズムを用いた強化学習を利用する方法は，推薦システムを取り巻く環境を学習する上で複数の課題に直面している [3]．本研究では，多種多様な環境で，逐次的に変化するユーザの行動パターンを学習するために，Soft Actor-Critic（SAC）アルゴリズムを利用したマルチタスク系列推薦フレームワークを提案する．

2 関連研究

強化学習を用いたマルチタスク系列推薦の既存フレームワークとして， Reinforcement Learning enhanced Multi-Task Learning (RMTL) がある [4]．RMTL の枠組みの中で，ユーザの行動パターンの理解を深めるために，AC アルゴリズムの一種である Twin Delayed Deep Deterministic Policy Gradient (TD3) アルゴリズム [5] が採用されている．TD3 アルゴリズムは，二種類の Critic を用いることで，それぞれの Critic ネットワークのうち，損失が最も少ないネットワークを選択し，学習プロセスを最適化する特徴を持っている．しかし，TD3 アルゴリズムは推薦環境におけるさまざまな行動の学習において一定の制約を有する [6].

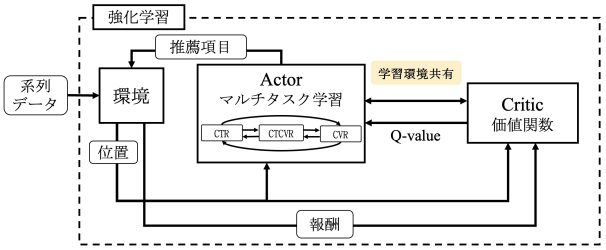

図 1　本研究の提案手法の概要図：情報共有を行う SAC アルゴリズムを用いたマルチタスク系列推薦フレームワーク

一方，AC アルゴリズムと比較してサンプリング効率が高く，より安定した学習が行える Soft AC (SAC) アルゴリズム [7] も存在している．SAC アルゴリズムは，ある状態に対する次の行動を確率分布で表現する環境で，未来の行動が現在の戦略に関係なく最適な行動を選択する off-policy 学習[*1]を採用することで，高効率なサンプリングを実現するアルゴリズムである．また，AC アルゴリズムの目的関数にエントロピ最大化項を考慮することで強化学習モデルの学習を安定化させる．SAC アルゴリズムは TD3 アルゴリズムほど強い制約が課せられていないため，SAC アルゴリズムを適用したマルチタスク系列推薦フレームワークは RMTL より汎用的に利用できる可能性がある．

3 提案手法

2 節で述べた強化学習を用いた汎用的な推薦システムの実現のために，本研究では図 1 の SAC アルゴリズムに基づくマルチタスク系列推薦フレームワークを提案する．

A Sequential Recommendation Framework using Soft Actor-Critic
† HONG Hyejin, Faculty of Culture and Information Science, Doshisha University
‡ KIMURA Yusuke, Graduate School of Culture and Information Science, Doshisha University
§ HATANO Kenji, Faculty of Culture and Information Science, Doshisha University

[*1] 逆に，未来の行動が現在の戦略に則って行動を選択する学習を on-policy 学習という．

SAC アルゴリズムをマルチタスク系列推薦フレームワークに適用する際には，学習過程で推薦項目の選択が過剰に評価される可能性があり，推薦の効率性や正確性に大きな影響を与える恐れがある [3]．推薦項目の選択が過剰評価される問題に対処するため，推薦項目を選択する環境とその選択に対応する価値を共有する機能を本研究の提案フレームワーク内に組み込むことで，過剰評価を防ぐことを目指す．

また，推薦システムでは複数の予測タスクを利用する必要があるため，Actor 部分でマルチタスク学習を行う必要もある．マルチタスク学習はそれぞれのタスクが異なる目標を持ちながらも，タスクに共通する特徴やパターンを学習することを可能にした学習法である．推薦システムの場合，どの商品がクリックされるかを予測する Click-Through Rate (CTR) やどの商品が購入されるかを予測する Conversion Rate (CVR)，クリック後に商品を購入につながる Click-Through Conversion Rate (CTCVR) などが予測タスクとなる．

4 評価実験

提案手法の性能を評価するために，RMTL と提案手法で推薦性能損失（a-loss と q-loss の和，a-loss: Actor で発生する損失，q-loss: Critic で発生する損失）を比較する．また，SAC アルゴリズムの影響を確認するために，RMTL のマルチタスク学習の環境を本研究の提案手法と同様の LS (Linear Scalarization) にした RMTL-LS との比較実験も行う．

評価実験で使用したデータセットは，EC サイトのユーザ行動履歴（クリック，購入履歴等）で構成されている Retailrocket recommender system dataset[*2] である．また，マルチタスク学習を行うためのモデルには RMTL の評価実験に用いられた CTR 予測に特化したモデルである Entire Space Multi-Task Model（ESMM）[8] を本研究の提案手法と RMTL-LS，RMTL で利用する．

その結果，提案手法の推薦性能が最も高いことが判明した（表 1 参照）．また，a loss，q loss 共に RMTL と RMTL-LS より提案手法の方が低かったことから，フレームワークを TD3 アルゴリズムで実現するよりも， 学習環境を共有する SAC アルゴリズムで実現する有効性が確認できた．

本研究の提案手法で推薦性能が向上した理由として，二つの貢献が考えられる．まず，ある状態に対する次の行動を確率分布で表現する環境でより多くのユーザの状況を学習可能な off-policy を用いた SAC アルゴリズムを利用したことで，意図的に以前と異なる行動を積極的に選択させ，TD3 アルゴリズムよりさまざまな推薦環境に対する学習ができた．また，学習環境を共有する本研究の提案手法により過剰評価が改善でき，q-loss の減少に繋がったと考えられる．

表 1　評価実験の結果：太字は各評価指標で最も性能が良い手法の結果を表す

	RMTL	RMTL-LS	提案手法
a-loss	2.753	2.740	**1.577**
q-loss	1.000	0.996	**0.235**
推薦性能損失	3.753	3.736	**1.812**

5 おわりに

本研究では，SAC アルゴリズムを適用したマルチタスク系列推薦フレームワークを提案した．本研究の提案手法は，State-of-the-art である RMTL より推薦性能が高いことを評価実験から確認した．

しかし，オフライン学習で評価可能なデータセットは，本研究の評価実験で使用した一種類のみであるため，他の性能検証方法である A/B テストを実行することで，構築したモデルの性能を精査する必要がある．

謝辞

本研究は日本学術振興会科学研究費助成事業基盤研究（B）23H03694，21H03555，および同志社大学文化情報学部学部・奨励学生制度の助成を受けて遂行されたものである．

参考文献

[1] Yuanguo Lin et al. A Survey on Reinforcement Learning for Recommender Systems. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–21, 2021.
[2] Timothy P. Lillicrap et al. Continuous control with deep reinforcement learning. arXiv:1509.02971, 2019.
[3] Xin Xin et al. Self-supervised reinforcement learning for recommender systems. page 931 – 940, 2020.
[4] Ziru Liu et al. Multi-Task Recommendations with Reinforcement Learning. In *Proceedings of the ACM Web Conference 2023*, page 1273 – 1282. ACM, 2023.
[5] Scott Fujimoto et al. Addressing Function Approximation Error in Actor-Critic Methods. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 2018.
[6] Sindhuja Bangari et al. A Review on Reinforcement Learning based News Recommendation Systems and its challenges. pages 260–265, 2021.
[7] Tuomas Haarnoja et al. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, 2018.
[8] Xiao Ma et al. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, page 1137 – 1140. ACM, 2018.

[*2] Retailrocket recommender system dataset, `https://www.kaggle.com/datasets/retailrocket/ecommerce-dataset`, アクセス日：2024 年 1 月 11 日