Adaptive Recommendation System Strategies: An Exploration of Online Machine Learning Algorithms

Fenglin Lu
Aberdeen Institute of Data Science and Artificial Intelligence, South China Normal University, Foshan, Guangdong, China

Keywords: MAB, Reinforcement Learning, Adaptive Recommendation System.

Abstract: This paper investigates the application of online machine learning algorithms in adaptive recommendation systems, focusing on the integration of reinforcement learning and multi-armed bandit (MAB) frameworks to enhance user experience and platform efficiency. As digital environments become increasingly dynamic, traditional recommendation systems face challenges such as the exploration-exploitation dilemma and the cold start problem. To address these issues, adaptive strategies that leverage the real-time decision-making capabilities of reinforcement learning and the optimization potential of MAB algorithms are explored. The research demonstrates how these technologies improve personalization and operational efficiency by dynamically adjusting to user preferences and behaviors. This paper serves as a valuable resource for individuals seeking to comprehend and explore the utilization of MAB algorithms in recommendation systems, while also shedding light on potential avenues for future research in the domain of online recommendation systems.

1 INTRODUCTION

Recommendation systems have become integral to our lives as a byproduct of the rapid development of information technology. According to Chen (2002), with the fast-paced advancement of online and mobile technologies, the internet, while providing a wealth of information to users, also leads to problems caused by information overload. This refers to the massive amount of information that prevents users from quickly and accurately pinpointing the information they need. Therefore, timely capturing user preferences, delivering content of interest, and enhancing user satisfaction with recommendation systems are crucial. These factors drive substantial traffic to recommendation systems, creating a positive feedback loop. Personalized recommendation technology is a key method to address information overload, offering users appropriate recommendation lists to improve satisfaction. In this process, recommendation systems continuously collect and store user attribute information, manage item feature information, and record user interaction behaviors to provide the most satisfactory recommendation list possible. Additionally, personalized recommendation systems can adapt to changes in user interests and preferences, significantly enhancing user experience.

However, the application of personalized recommendation systems faces two major challenges: the exploration-exploitation dilemma and the cold start problem. When facing a new user, due to the lack of historical data, it is challenging to effectively recommend products, resulting in suboptimal performance. The introduction of any new user or product experiences a transition from no data to rich data, and effectively resolving the data deficiency at the initial stage is known as the "cold start problem (Jiang, 2024)." Over the long term, the recommendation system must balance whether to frequently recommend the best-known items to users or to explore different items for potentially better outcomes. The multi-armed bandit (MAB) algorithms are specifically designed to allow recommendation systems to make adaptive choices, maximizing returns and optimal feedback through each selection. Common recommendation systems include various online news websites, video platforms, and short video apps on mobile devices. These platforms collect user preference information based on browsing history to recommend videos or other content that might interest the user.

Lu, F.
Adaptive Recommendation System Strategies: An Exploration of Online Machine Learning Algorithms.
DOI: 10.5220/0012949200004508

In *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence (EMITI 2024)*, pages 447-451
ISBN: 978-989-758-713-9

2 INTRODUCTION TO REINFORCEMENT LEARNING THEORY

In the field of machine learning, traditional learning methods are mainly divided into two types: supervised learning and unsupervised learning. Although powerful, these methods have limitations when dealing with dynamic decision-making problems in practical applications. In particular, they cannot actively select training samples or dynamically adjust strategies based on the results of previous actions. Unlike these methods, reinforcement learning offers a distinct learning paradigm focused on sequential decision-making problems. It enables learners not only to make choices but also to adjust their future actions based on the outcomes of these choices, aiming to maximize long-term objectives.

2.1 Overview of Reinforcement Learning

Reinforcement learning is about learning behavior and decision-making, where an agent needs to make a series of decisions through interaction with the environment. In this process, the agent takes actions and receives feedback from the environment, typically in the form of rewards, with the goal of maximizing cumulative rewards. This learning mode makes reinforcement learning particularly suitable for applications that require continuous decision-making, such as games, robot control, and online recommendation systems(Ke et al., 2023).

In recommendation systems, the interaction between users and the recommendation system is inherently sequential. Recommending the best items is not just a prediction problem but a complex multi-step decision-making problem. This makes the recommendation problem very suitable for modeling through reinforcement learning, where the recommendation algorithm continuously learns how to adjust its recommendation strategy based on user feedback to enhance user satisfaction and overall system performance. In this context, the recommendation algorithm acts as a reinforcement learning agent, with the rest, including users and items, constituting the environment of the agent. This is what constitutes a Markov decision process.

2.2 Markov Decision Process

The Markov Decision Process (MDP) is a fundamental framework of reinforcement learning and a primary research direction for addressing sequential decision-making issues. It is mainly used to solve optimal decision-making problems in stochastic dynamic systems. Markov properties, which imply that the next state transition depends solely on the current state and not on the sequence of events that preceded it, determine future state transitions along with current actions and states. The main elements of a Markov decision process include the decision period (T), system states (S), action sets (A), state transition probabilities (P), and reward functions (R). Each MDP can be represented using these five elements.

1. Decision Epochs (T): The decision-making process can be categorized based on the length of decision epochs into finite and infinite decision processes. Depending on whether the decisions are made at discrete intervals or continuously, it can be further divided into discrete and continuous decision problems. For discrete finite Markov decision processes, the decision epoch refers to the time period between the start and the end of a decision.

2. State Space (S): The set of all possible states at any decision point within the system, where each state is characterized by its unique properties.

3. Action Space (A): For any state $s \in S$ at any decision point, the set of all possible actions that can be taken is defined as the action space.

4. State Transition Probabilities (P): For a given state and action, the system state s transitions to another state s_{t+1} with a certain probability

$$p_{ss'}^{a} = p(s_{t+1} = s' \mid s_t = s, a_t = a) \quad (1)$$

The transition probabilities, along with the current state and action, determine the next state of the system.

5. Reward Function (R): In an MDP, the reward function is an expectation indicating the reward received when taking an action in a particular state. Additionally, a discount factor is defined to discount future rewards, which can be described in terms of costs, profits, or other benefits.

3 OFFLINE RECOMMENDATIONS AND ONLINE LEARNING

3.1 Offline Recommendations

The process of reinforcement learning includes iterative collection of experience through interaction with the environment, usually using the latest learning strategies, and then using these experiences to improve strategies. However, in the early stages of development, due to the limitations of computer power, real-time interaction was somewhat impractical. Given this context, offline learning has garnered attention in the field of recommendation systems. Offline recommendation does not require interaction with the environment during the learning process, significantly reducing the computational power requirements. Offline recommendation systems use historical offline data from users to learn the corresponding static recommendation models. Typically, under a given set of users, items, and a certain amount of user-item rating matrix, various methods are used to estimate the unknown ratings in the rating matrix, and then recommend items with high ratings to users(Wang, 2021). The ratings are divided into implicit and explicit ratings. Implicit ratings are based on user behaviors that reflect their preferences, such as clicks, bookmarks, browsing duration, etc. Explicit ratings are much more straightforward, based on user ratings of items that indicate their preference level, such as star ratings for movies, satisfaction ratings for products, etc., usually presented in a certain numerical range, with higher numbers indicating greater user preference.

3.2 Online Learning and Real-Time Recommendations

While offline learning reduces the threshold for computational power, it often does not achieve ideal results in highly dynamic fields, such as news and advertising recommendations. One significant reason is that offline learning is somewhat powerless against the cold start problem and cannot promptly track changes in user preferences. As computer performance improves and user experience demands further increase, the need for research into online learning in the field of recommendation systems has gradually expanded. Due to the fluidity of user preferences, recommendation systems must not only focus on users' current interests but also explore new content that users might find interesting. This is the previously mentioned exploration-exploitation problem. Reinforcement learning adapts well to such issues, and the multi-armed bandit algorithm, as a simple reinforcement learning method, utilizes the sequential interaction characteristics of reinforcement learning while avoiding the complexity and computational intensity of other reinforcement learning algorithms. It is an excellent way to handle online recommendation issues. Online learning, a subclass of reinforcement learning, can effectively overcome the drawbacks of traditional batch learning. When new data arrives, the model can be updated immediately, offering timely effectiveness(Liu, 2022). The bandit algorithm belongs to partially feedback-based online learning.

4 REINFORCEMENT LEARNING AND THE MULTI-ARMED BANDIT PROBLEM

The Multi-Armed Bandit (MAB) problem is a classic issue in reinforcement learning, focusing on how to make the best decisions among a series of choices to maximize the total return. This problem originates from the gambling world, where a gambler faces multiple slot machines (each representing an "arm"), each with different payout rates. The gambler's goal is to maximize the total reward through a series of lever pulls, deciding which machine to play, how often to play each machine, in what order, and whether to continue with the current machine or try different ones. Addressing this problem involves a crucial dilemma, previously mentioned as the exploration-exploitation problem(Sun, 2023). The decision is whether to pull the arm with the highest expected return or to explore other machines to gather more information about potential returns.

In sectors like recommendation systems, medical treatment allocation, and financial investments, the MAB problem provides a framework to help algorithms balance exploration (seeking more information to improve future decisions) and exploitation (using current knowledge to maximize immediate returns). This balance is achieved through different strategies. One approach widely adopted is the ε -greedy algorithm, which balances between exploring new options and exploiting known preferences. This algorithm initially explores with a probability ε and selects the best-known option with a probability of 1- ε . Although simple, the ε -greedy algorithm can effectively improve recommendation performance in dynamic environments.

Another prominent algorithm is the Upper Confidence Bound (UCB) method, which calculates an upper confidence bound for each arm's expected reward and selects the arm with the highest upper confidence bound a_t. The formula for the confidence bound is as follows:

$$UCB(a_t) = \hat{\mu}_a + \sqrt{\frac{2\ln t}{n_a}} \quad (2)$$

where $\hat{\mu}_a$ represents the average reward (expected return) of action a, n_a is the number of times action a has been chosen, and t is the current total number of actions taken. The square root term represents the exploration factor, which decreases as the number of times the action is chosen increases, indicating a preference for actions that have not been explored frequently. This naturally balances exploration and exploitation.

Thompson Sampling is a probabilistic algorithm that adjusts the Beta distribution for each action a after each time step, employing the sampled value $\theta_a(t)$ from that distribution. The Beta distribution parameters for an action a at time t are updated to

$$\text{Beta}(S_a(t) + 1, F_a(t) + 1) \quad (3)$$

where $S_a(t)$ denotes the number of successes and $F_a(t)$ the number of failures of action a up to time t. This $+1$ approach to both success and failure counts ensures a non-zero probability, maintaining a smoother distribution even in the absence of any trials or evidence. The algorithm progressively learns and approaches the true success probability of each action by updating the Beta distribution parameters of each action (arm). It effectively handles uncertainty and adapts to changes in the environment.

5 APPLICATION OF MULTI-ARMED BANDITS IN RECOMMENDATION SYSTEMS

The study of the MAB problem is not limited to theoretical models but is also extensively applied in the design and optimization of algorithms in practical applications. For instance, in online advertising, MAB algorithms can determine which ad layout is most likely to attract user clicks, while in recommendation systems, they help the system decide when and what content to recommend to a specific user to maximize user satisfaction and platform revenue. These algorithms are also used in dynamic pricing, resource allocation, and other decision-making areas.

5.1 e-Commerce Platform Product Recommendation

On e-commerce platforms, as the variety of products increases and the market environment continuously evolves, product iteration becomes more frequent. Although the market has long been dominated by best-selling items, niche products, also known as non-bestsellers, are gradually revealing their unique market value. Although these products do not have high individual sales, their diverse types can aggregate into significant volumes, known as the long tail phenomenon. The long tail effect emphasizes "personalization," "customer power," and "small profits but large market(Zhang, 2022)." In the Internet era, online stores are not limited by physical space and can store and display a large number of niche products. Coupled with technologies such as search engines, consumers can more easily find and purchase these products that are difficult to obtain through traditional channels. This market dynamic poses new challenges for recommendation systems: how to recommend those few potentially interesting long-tail products among a vast array of items.

The MAB algorithm can assist recommendation systems in selecting items to display to users, achieving a balance between exploring new items and exploiting popular ones through continuous learning and adjustment. This helps promote the sale of popular products while enhancing the exposure and recommendation of long-tail items, thereby increasing overall sales revenue(Cesa-Bianchi & Lugosi, 2012). Zeng et al. designed a Thompson Sampling-based MAB algorithm, which, combined with users' historical behavior and item metadata, significantly improved the effectiveness of recommendations.

5.2 News Article Recommendation

On news websites, each article can be considered an arm, with the click-through rate as the reward. For example, the Yahoo! homepage news recommendation system uses a Multi-Armed Bandit (MAB) framework based on the Upper Confidence Bound (UCB) algorithm. By dynamically adjusting the display strategy of articles, the system can balance between the timeliness of news and user interests. This improvement has increased the overall click-through rate by 12.5%(Li et al., 2010). This example

highlights the effectiveness of MAB algorithms in responding to user behavior in real time.

5.3 Music/Video Recommendation

On music or video streaming platforms, each song or video can be considered an arm, with user's play completion rate or like rate as the reward. By applying context-aware MAB algorithms, music recommendation systems can integrate various contextual factors such as the user's current mood, activities, and more. This enables more accurate predictions of the user's music preferences, leading to the generation of more personalized and dynamic recommendation lists. Consequently, it enhances user satisfaction with the platform and increases their usage duration(Wang et al., 2014).

5.4 Online Advertising Placement

In the online advertising field, each advertisement can be seen as an arm, with the click-through rate (CTR) or conversion rate (CVR) as the reward. Research by Yang (2018) has demonstrated the effectiveness of Thompson Sampling in predicting the value of advertising opportunities, thereby aiding advertisers in increasing overall revenue. This highlights the efficiency of MAB algorithms in optimizing the allocation of advertising resources.

6 CONCLUSIONS

This thesis has explored the intricate role of online machine learning algorithms within the context of adaptive recommendation systems. Through the detailed study of reinforcement learning and multi-armed bandit (MAB) problems, it has been demonstrated how these techniques significantly enhance the performance and adaptability of recommendation systems across various domains, including e-commerce, news aggregation, music/video streaming, and online advertising. The implementation of MAB algorithms has addressed the critical exploration-exploitation dilemma effectively. By optimizing the selection process through algorithms such as Ɛ-greedy, UCB, and Thompson Sampling, recommendation systems can balance between exploring new options and exploiting known user preferences, thereby improving user engagement and platform profitability. The application of these advanced algorithms has led to a more personalized user experience, as systems can offer recommendations that align closely with individual user preferences and behavioral patterns. A key challenge of the MAB problem is designing strategies that can quickly adapt to environmental changes and achieve optimal performance over the long term. This requires thorough theoretical analysis and experimental verification of different strategies to ensure they perform well in various scenarios. Additionally, since online recommendation systems heavily rely on user data to operate, how to adequately protect user privacy is also a direction that needs further research.

REFERENCES

Chen, K. (2022). Research on personalized learning systems based on multi-armed bandit algorithms (Master's thesis, Nanjing University of Posts and Telecommunications).

Jiang, F. (2024). Research on cold-start recommendation algorithms based on meta-contrastive learning (Master's thesis, Beijing University of Posts and Telecommunications).

Ke, K., Jin, S., Gao, B., & Huang, X. (2023). Robot multi-contact interaction task control based on reinforcement learning. *Journal of Dynamics and Control*, (12), 53-69.

Wang, R. (2021). Research and implementation of offline recommendation algorithms based on user review data (Master's thesis, Southwest Jiaotong University).

Liu, F. (2022). Research on personalized recommendation methods based on user behavior sequence mining (Doctoral dissertation, Harbin Institute of Technology).

Sun, Y. (2023). Portfolio strategies based on multi-factor models and multi-armed bandit algorithms (Master's thesis, Shandong University).

Zhang, Y. (2022). Dynamic pricing algorithms for niche products based on the MAB model (Master's thesis, Nanjing University)

Cesa-Bianchi, N., & Lugosi, G. (2012). Combinatorial bandits. *Journal of Computer and System Sciences, 78*(5), 1404-1422.

Zeng, C., Wang, Q., Mokhtari, S., & Li, T. (2016, August). Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 2025-2034).

Li, L., Chu, W., Langford, J., & Schapire, R. E. (2010, April). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International conf. on World wide web* (pp. 661-670).

Wang, X., Wang, Y., Hsu, D., & Wang, Y. (2014, July). Exploration in interactive personalized music recommendation: a reinforcement learning approach. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 11*(1), 1-22.

Yang, C. H. (2018). Real-time price prediction model for advertising based on Thompson Sampling and truncated regression (Master's thesis, South China University of Technology).