A Analytical and Practical Insights into Multi-Armed Bandit Problems in Recommendation Systems

Maike Feng *

School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Shenzhen, 518172, China

* Corresponding Author Email: 222012027@link.cuhk.edu.cn

Abstract. This paper delves into the application of the Multi-Armed Bandit (MAB) algorithm in recommendation systems, a tool increasingly prevalent across diverse sectors such as e-commerce, social networks, and news platforms. The primary objective of these systems is to curate content that resonates with user preferences, thereby enhancing user engagement and augmenting business revenue. Central to the optimization of these recommendation strategies is the careful balance between exploration - the pursuit of new, potentially relevant options - and exploitation - the utilization of known, popular choices. The MAB algorithm, an online learning method, adeptly navigates this balance. This study presents a detailed exploration of the MAB algorithm's theoretical underpinnings and its practical applications in recommendation systems. We implement these concepts using real-world datasets to assess their efficacy in such systems. The paper concludes by examining the benefits and constraints of employing MAB algorithms in recommendation contexts and proposes avenues for future research. This analysis aims to contribute to the ongoing evolution of recommendation systems, underscoring the pivotal role of MAB algorithms in their advancement.

Keywords: Multi-Armed Bandit problem; Recommendation Systems.

1. Introduction

Recommendation systems have become a cornerstone of the modern information society, with pivotal roles in e-commerce, social media, and news consumption [1]. Their primary aim is to deliver personalized recommendations, curating content that aligns closely with user preferences to enhance user satisfaction and drive business revenue. Crafting an effective recommendation system, however, presents significant challenges. A key hurdle is balancing exploration and exploitation. Exploration involves recommending novel content to users to gather more insights about their preferences; exploitation, conversely, focuses on suggesting content that is likely to be well-received based on existing user data. Striking the right balance is crucial: excessive exploration can inundate users with irrelevant content, degrading their experience, while excessive exploitation may restrict their exposure to new and potentially interesting content.

This paper explores the application of the Multi-Armed Bandit algorithm as a solution to this exploration-exploitation dilemma in recommendation systems. We aim to provide a comprehensive theoretical foundation and practical insights for the development and optimization of these systems.

The paper is structured as follows: Section 2 introduces the MAB problem, its algorithms, and their application in recommendation systems; Section 3 assesses the performance of these algorithms using real-world datasets; Section 4 discusses the advantages and limitations of the MAB approach. Section 5 concludes with a summary of the key findings and contributions of this study.

2. Literature Review

The Multi-Armed Bandit problem, also called as K-armed bandit problem, a classic problem in reinforcement learning, is a simple model for the exploration vs. exploitation trade-off [2, 3]. This problem is originated in the gambling industry, where "multi-armed" refers to the tables in a casino and "bandits" are the games at each table. The problem can be described as a sequential decision model to maximize the cumulative reward based on the previous information. The player needs to choose an arm (action) among the set of arms $\mathcal{A}$ at each trial and the selection will result in a certain

reward R. According to different rewards of arms, one possible policy is exploration – pulling different arms to get more information and another is exploitation – pulling the arm with the best reward in the past. Thus how to deal with the trade-off is the key of MAB problem. Classic solutions, such as Ɛ-Greedy , Upper Confidence Bounds(UCB) and Thompson Sampling(TS) handle this goal in distinct ways [4-6].

2.1. Algorithms in Multi-Armed Bandit Problems

2.1.1. Ɛ-Greedy

The Ɛ-Greedy algorithm handles the exploration-exploitation trade-off by randomly selecting an arm to maximize the cumulative rewards. The algorithm selects an exploration arm with probability Ɛ and the current best arm, defined by formula (1), with probability 1- Ɛ.

$$a_t^* = argmax_{a\in\mathcal{A}} R_t(a) = argmax_{a\in\mathcal{A}} \frac{1}{N_t(a)} \sum_{\tau=1}^{t-1} r_\tau \tag{1}$$

The advantage of the policy is making the most of current information while keeping exploration. Because of its effectiveness and efficient, this algorithm is widely use in both academic experiment and practice and can usually delivers robust result. However, the key challenge is to define an appropriate Ɛ. If it is too large, the algorithm may explore too much and waste a lot of known information. In turn, if it is too small, the algorithm may be too conservative to fully explore possible options. Therefore, a balance need to be find to make the algorithm not only obtain high payoff but carry out exploration under the condition of controllable risks.

2.1.2. Upper Confidence Bounds

The key of UCB algorithm is to balance exploration and exploitation through the upper confidence bounds. In each trail, the algorithm takes both average rewards and uncertainty of each action into consideration. Specifically, the algorithm will calculate the upper confidence bound for each arm based on the average reward and standard deviation of it. And at each trail, the arm with highest upper confidence bound will be selected to maximize expected reward:

$$a_t^* = \ argmax_{a\in\mathcal{A}} \ R_t(a) + \sqrt{\frac{2\log t}{N_t(a)}} \tag{2}$$

Where $N_t(a)$ means the number arm a to be chosen until trial t.

In this way, UCB algorithm would balance exploration and exploitation of each arm according to their uncertainty. The arms with large uncertainty would have lower upper confidence bounds and have lower probability to be selected. On the contrary, the arm with small uncertainty and high average reward would have higher upper confidence bounds and more likely to be selected. Moreover, another advantage of UCB algorithm is the ability to deal with the correlation of different arms. In MAB problems, the interaction between different arms may result in better rewards of some arms than others. And UCB algorithm can adjust the upper confidence bounds according to the correlation of different arms in order to balance exploration and exploitation better.

2.1.3. Thompson Sampling (TS)

Thompson Sampling algorithm, the concept of which was proposed in 1933 by William R. Thompson and implemented in 2011, is a sampling method based on Bayesian theory [7, 8]. It predicts the probability distribution of each arm by sampling, and selects arms based on the distribution. To be specific, TS algorithm builds a prior distribution (beta distribution generally) for each arm and then sample randomly according to the distribution of each arm. At each trial, action a is selected according to the probability that a is optimal according to the history of actions h_t already known by the player:

$$\pi(a|h_t) = \mathbb{P}[R(a) > R(a'), \forall a' \neq a|h_t] = \mathbb{E}[a = \ argmax_{a\in\mathcal{A}} \ R(a)] \tag{3}$$

TS algorithm takes into account the different probability of rewards for different arms and has the probability of denying the arms with larger current expected rewards and selecting arms with less current expectation. Besides, the algorithm also has advantages like simplicity and online learning and thus widely used in advertising, recommendation system and other scenarios.

2.1.4. Gradient Bandit Algorithm

In Gradient Bandit Algorithm, each arm is connected with a 'preference' value according to which to choose an arm at each trial. The algorithm updates 'preference' by calculating gradient of the preference of each arm which is often defined as a function about the arm, the output of which reflects the expected reward of the arm. When calculating the gradient, the algorithm considers both the known preference and balance between exploration and exploitation to determine the probability of choosing each arm. In reality, softmax policy is usually used to choose arms (Formula (4)) and updates the preference based on the actual reward: If the actual reward of the arm larger than expected reward, the preference increases; if the actual reward is less than the expectation, the preference decreases [10]. The algorithm has good theoretical property and practical application effect and is widely used in recommendation systems nowadays.

$$\pi_t(a) = \frac{e^{H_t(a)}}{\sum_{a' \in \mathcal{A}} e^{H_t(a')}} \tag{4}$$

Where $\pi_t(a)$ is the probability that arm a is selected in trial t and $H_t(a)$ is the 'preference' of the player to arm a in trial t

2.2. Application of Multi-Armed Bandit Problems in Recommendation Systems

The main objective of recommendation system is to offer users the most relevant and interesting content. That means it has to continually choose an item and receive users' feedback and update itself according to this which is similar with the multi-armed bandit where each item can be seen as an arm in MAB problem. Thus, many multi-armed bandit algorithms can be well applied in recommendation system in different scenarios.

Bayesian bandit usually models the users' interests and predict the expected payoff of different recommendation items using this model. At each trial, recommendation item with the highest posterior probability and expected reward is selected. This method is suitable for scenarios that have some prior knowledge and can model users' interests well. UCB algorithm in recommendation systems calculates the upper confidence bounds of items to choose the best optimal items instead of modeling users' interests. Thus it is suitable for the scenarios where users' interests change frequently and hard to model accurately. And in gradient bandit algorithm, to maximize user satisfaction, the probability of each arm is iteratively adjusted in recommendation systems. This kind of bandit is widely used in where the number of recommendations is large and users' interests are continuous. In general, many recommendation systems can be modeled as MAB problems in practical application. In the next section, the performance of MAB algorithms in recommendation system would be evaluated by experiment using real data set.

3. Experimental Evaluation

3.1. Experimental Setup

In order to evaluate the performance of MAB algorithm in recommendation systems, an experiment has been done using the MovieLens dataset. The MovieLens dataset is one of the most commonly used benchmark datasets in the field of recommendation systems and is widely used to research and test various recommendation algorithms. These datasets were collected over various periods of time by the GroupLens Lab of University of Minnesota. MovieLens Latest Datasets were used in this paper, which consist of users' ratings to different movies, movie information, tags of movie and link information.

In this experiment, movies are modeled as the arms in MAB problems and users' ratings to movies are the rewards of different movies. Three main MAB algorithms mentioned in 2.2 (Bayesian Bandit, UCB bandit, Gradient Bandit) were implemented in the experiment to recommend movies to users. In Bayesian bandit, Thompson method was used to sample from the distribution of each arm (movie). In UCB bandit, to avoid the denominator in formula (2) becomes zero, each movie would be selected once to initialize the ucb value in advance. Besides, the hyperparameter of the step size α in gradient bandit in this experiment was set to be 0.1.

To measure the performance of the recommendation results of the MAB algorithms, recommendation accuracy is defined as the ratio of the number of movies accepted by users to the total recommendation. And in the experiment, if the movie recommended has higher rating, it is thought to be more likely to be accepted by users. Moreover, to compare the performance of MAB algorithms and traditional recommendation algorithms, item-based collaborative filtering method was used as a baseline.

3.2. Result Analysis

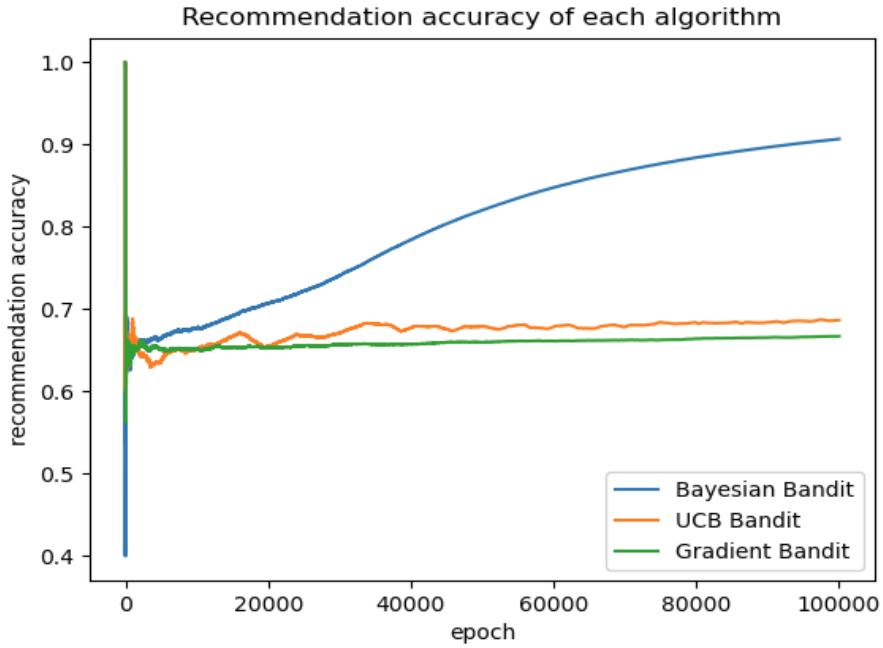

Fig 1. Recommendation accuracy of three MAB algorithms (Photo/Picture credit: Original).

The performance of each MAB algorithm is plotted in Fig 1. It shows that the accuracy of gradient bandit reaches convergence very quickly and UCB converges after a period of slight fluctuation while Bayesian keeps increasing after initial fluctuation. The performance of UCB is a little better than Gradient bandit but they neither exceed 70%. According to the trend of Bayesian, the accuracy of it would stabilize at around 90%. The reason of the obvious gap is that it is supposed that users' interest does not change and the prior knowledge of the ratings can improve the performance of Bayesian as iterations increase.

Table 1. Performance of MAB algorithms and traditional recommendation algorithm.

Algorithms	Recommendation Accuracy
Item-based collaborative filtering	0.3731
Bayesian Bandit	0.9069
UCB Bandit	0.6867
Gradient Bandit	0.6672

The final performance after 100,000 iterations of each kind of bandit is also recorded in Table 1 to compare with the traditional recommendation algorithm. It is obvious that the performance of MAB algorithms in movie recommendation system is better than the traditional recommendation systems.

4. Discussion

Although the application of multi-arm gambling machine algorithms in recommendation systems has achieved some positive results, there are still many limitations to be overcome. This section will provide an in-depth analysis of the advantages and limitations of MAB algorithms.

First of all, MAB algorithm is a kind of online learning algorithm. It can collect the latest feedback information in real time and update the model immediately in order to adapt to the changes of the environment better [11]. Thus it has significant advantages when dealing with dynamic changing environment. Besides, because of the balance of MAB algorithm between exploitation and exploration, it can deal with the cold-start problem in recommendation systems effectively [12]. What is more, MAB algorithm does not only consider global optimization but also attaches importance individual feedback thus can offer more personalized recommendation service [13].

However, the computation of MAB has always been a big problem, especially facing large-scale problems. That is because MAB needs to calculate the expected reward of every arm and the computation would be extremely complex if there are too many arms. Another challenge of it is that the performance of MAB algorithm depends on the quality of data. As the algorithm updates the model through users' feedback, the noise in the feedback data and the frequent changes of users' behavior pattern would affect the performance of the algorithm. To overcome these challenges, there are many ways to explore. For example, reduction of computational complexity can be implemented by parallel computation or optimization algorithms; introducing noise processing mechanism or designing a more stable model can also reduce the dependence on data quality to a certain extent.

5. Conclusion

This paper presents a comprehensive review of the fundamental principles, classifications, and principal applications of the Multi-Armed Bandit algorithm, with a focus on its deployment and performance assessment within recommendation systems. Our findings reveal that the MAB algorithm offers substantial benefits in navigating specific challenges encountered in recommendation systems, such as balancing exploration and exploitation, addressing cold-start issues, and adapting to dynamic environmental shifts. Despite its strengths, the MAB algorithm is not without limitations, including substantial computational demands and a strong reliance on the quality of data. Nevertheless, the inherent advantages of the MAB algorithm suggest its considerable potential in recommendation system applications. Its effectiveness is particularly notable in large-scale and dynamically evolving recommendation contexts, where the algorithm's online learning capabilities and adaptability render it a highly promising solution. Looking ahead, there is potential for integrating the MAB algorithm with advanced models such as deep learning or even GPT architectures. These models excel in processing complex nonlinear relationships and extracting sophisticated features, which could mitigate the computational challenges of MAB and further enhance the overall system performance.

Reference

[1] Ricci, F., Rokach, L., & Shapira, B. (2010). Introduction to recommender systems handbook. In Recommender systems handbook (pp. 1-35). Boston, MA: springer US.

[2] Zhao, X., **a, L., Tang, J., & Yin, D. (2019). "Deep reinforcement learning for search, recommendation, and online advertising: a survey" by **angyu Zhao, Long **a, Jiliang Tang, and Dawei Yin with Martin Vesely as coordinator. ACM sigweb newsletter, 2019(Spring), 1-15.

[3] Varaiya, P., & Walrand, J. C. (1983, September). Multi-armed bandit problems and resource sharing systems. In Proceedings of the International Workshop on Computer Performance and Reliability (pp. 181-196).

[4] Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. Machine learning, 47, 235-256.

[5] Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 3(Nov), 397-422.

[6] Chapelle, O., & Li, L. (2011). An empirical evaluation of thompson sampling. Advances in neural information processing systems, 24.

[7] Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Biometrika, 25(3-4), 285-294.

[8] Agrawal, S., & Goyal, N. (2012, June). Analysis of thompson sampling for the multi-armed bandit problem. In Conference on learning theory (pp. 39-1). JMLR Workshop and Conference Proceedings.

[9] Zhu, X., Xu, H., Zhao, Z., & others. (2021). an Environmental Intrusion Detection Technology Based on WiFi. Wireless Personal Communications, 119(2), 1425-1436.

[10] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction. MIT press.

[11] Yan, C., Han, H., Zhang, Y., Zhu, D., & Wan, Y. (2022). Dynamic clustering based contextual combinatorial multi-armed bandit for online recommendation. Knowledge-Based Systems, 257, 109927.

[12] Silva, N., Silva, T., Werneck, H., Rocha, L., & Pereira, A. (2023). User cold-start problem in multi-armed bandits: When the first recommendations guide the user's experience. ACM Transactions on Recommender Systems, 1(1), 1-24.

[13] Zhou, T., Wang, Y., Yan, L., & Tan, Y. (2023). Spoiled for Choice? Personalized Recommendation for Healthcare Decisions: A Multiarmed Bandit Approach. Information Systems Research.