Proceedings of CONF-MLA 2024 Workshop: Semantic Communication Based Complexity Scalable Image Transmission System for Resource Constrained Devices
DOI: 10.54254/2755-2721/83/2024GLG0076

Advanced MiniMax Optimality Strategy in Bandit Problems

Sai Zhang

City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong SAR, China

saizhang7-c@my.cityu.edu.hk

Abstract. Nowadays, with the development of the Internet, recommending personalized information for every user becomes economically valuable. How to properly recommend to users the content they prefer is important. Mostly, bandit strategies are useful for this problem. As one of the most time-saving algorithms, Minimax Optimal Strategy (MOSS) performs well in most data sets. However, the traditional MOSS algorithm applies a conservative strategy in order to minimize the cumulative regrets in the worst situations. This strategy will over-explore the bandit models and spend too much time exploring every arm at the initial stage. This means the MOSS strategy cannot fully take advantage of the known information and then the accuracy and efficiency will be lower. In real life, people have a lot of Internet information to look through and they will not stay on one website or one advertisement for a long time. So, designing an efficient algorithm that performs well in all situations and can make good use of the known knowledge to do quick exploitation becomes so important. Therefore, this research aims to develop an advanced MOSS algorithm that can properly explore each of the bandit arms and improve the performance of the exploitation phase. By achieving this goal, this research adds an enhancement factor on the confidence bound part to speed up the exploration process. The research bases on the MovieLens data set and the data set is divided into 18 categories (arms) according to its "genre" attribute. Then apply this new advanced MOSS strategy to the designed data set and compare its performance with normal MOSS and UCB algorithms. Results display that the advanced MOSS strategy's selection provides higher total rewards and lower average cumulative regrets than others with the advantage of time-saving. This shows the improvement of the advanced MOSS strategy in balancing the exploitation and exploration phases.

Keywords: bandit strategy, MOSS, over-exploitation, enhancement factor.

1. Introduction

With the development of the Internet, information technologies are crucial for attracting users' attention. Providing appropriate knowledge for each different user can make them interested in websites and products. Therefore, the recommendation system plays an important role in this process and bandit strategies are common methods to do this task. In bandit problems, decision-makers choose one arm each round and get a reward from it. The players aim to maximize the total rewards and choose the best arm with the highest rewards. Bandit strategies are used in this process to better choose arms in each round. This problem was first raised by Robbins and used in clinical trials area [1]. Nowadays, with the development of the Internet and artificial intelligence, recommendation systems are widely used in some Internet organizations, like Amazon, Yahoo and so on [2]. Bandit strategies play an important role in today's information age and are necessary to build recommendation systems. As one of the most

DOI: 10.54254/2755-2721/83/2024GLG0076

efficient bandit algorithms, the MOSS strategy was first proposed by Audibert and Bubeck in 2009 to eliminate a logarithmic factor in the upper confidence bound algorithm [3] and then improved by Degenne and Perchet in 2016 to introduce an anytime version [4]. This algorithm reduces the calculation cost and can generate almost the most optimal strategy at any time. It can return the lowest average regret compared with the upper confidence bound (UCB) algorithm. Recently, researchers have tried one kind of subroutine with regret upper bound to implement the MOSS strategy [5]. To better realize the advantage of the minimax policy, researchers also combine MOSS with the Bayes policy [6].

However, the modified MOSS algorithm is still not the ultimate strategy. The design of the MOSS strategy is too conservative so it cannot fully balance the exploration and exploitation phases. It still spends too much time on arms with high uncertainty in the late part of the algorithm and causes over exploration problem. An algorithm that can't be well applied efficiently is not useful to recommendation systems because there is a lot of new information moment by moment and users will not spend too much time on one specific website or application. Therefore, designing an advanced MOSS algorithm that can perform well and efficiently and balance the exploration-exploitation phases by utilizing the former known information is important for the development of artificial intelligence. This research relies on the MoiveLens data set and tests the performance of the advanced MOSS algorithm with an enhancement factor added in the upper confidence bound part. This aims to reduce the attention on uncertain arms and focus more on exploiting arms instead of exploring unknown arms as time goes on. This algorithm should generate lower average regrets and more accurate recommendations for movies that users are likely to prefer than normal MOSS and UCB algorithms.

2. Literature Review

As a traditional reinforcement learning problem, the bandit problem has been explored for a long time to get the maximal benefits after the exploitation-exploration process. The simplest method is the random algorithm which just selects one arm randomly each time. Although it needs little time and calculation cost, this algorithm cannot use any of the previous knowledge and therefore will gain very low rewards. To better utilize the previous information, the greedy algorithm has been introduced to exploit the arm that has the highest expected rewards in anterior rounds [7]. This strategy will choose the best arm so far and, in some rounds, it will randomly choose one of the arms to reduce selection bias. However, the greedy strategy will not explore uncertain arms and cannot balance the exploitation-exploration phase which may miss better choices at the beginning. The upper confidence bound (UCB) algorithm can solve this problem by adding one uncertain bound to better explore unknown arms [8]. With the confidence bound, this algorithm can equally give each arm chances to show their real expected rewards as much as possible. However, the UCB algorithm may increase unnecessary penalty costs because of the uncertain exploration process. It spends too many unnecessary rounds to ensure the accuracy of the model. The Thompson Sampling can optimize this process by introducing posterior probability [9]. Based on prior distribution and posterior probability, this strategy can select the best arm with few rounds. As one of the most popular algorithms, the Thompson Sampling strategy performs much better than other algorithms in most situations, but it depends too much on prior distribution and cannot make optimal decisions in the worst cases or complicated environments. The Minimax Optimal Strategy (MOSS) [10] is designed to make sure that the regret in the worst case also has good boundaries and it solves the over-exploration problem. Although the MOSS algorithm is almost the ultimate algorithm, it still has one main problem: it is too conservative and cannot well balance the exploitation and exploration phases. Therefore, this research paper aims to modify the traditional MOSS algorithm and try to overcome the over-exploration problem by adding an enhancement factor to decrease the influence of uncertain arms with the passage of time.

3. Methodology

UCB algorithm was the best strategy to constrain the confidence bound on bandit problems, but the development of the Minimax Optimal Strategy in the Stochastic case (MOSS) makes the regret bounds and variant more accurate [11]. The MOSS strategy serves the bandit problems better than UCB.

DOI: 10.54254/2755-2721/83/2024GLG0076

However, the normal MOSS strategy explores too much on every bandit's arm at the initial stage to minimize the worst situations. Therefore, this paper aims to modify the traditional MOSS strategy and hopes to reduce the exploration times of arms with uncertainty over time.

3.1. MOSS Strategy

We consider a multi-bandit arm set $v = (v_1, v_2, ...v_k)$ which has unknown reward distributions d_i for each arm in the set v [12]. There are k arms and n total action rounds. In the first k rounds, the player chooses each arm one by one to build up the original data information. After k rounds, the player should subsequently select the best arm following this formula [13]:

$$A_t = arg\max_i \mu_i(t-1) + \sqrt{\frac{4}{T_i(t-1)} log^+\left(\frac{n}{kT_i(t-1)}\right)} \quad (1)$$

Where $log^+(x) = log\, max\{1, x\}$; μ_i is the current average reward of arm i; $T_i(t-1)$ means the selected times of arm i until t-1 rounds.

This is the traditional MOSS algorithm. It combines the reward information selected in each round and an updated confidence bound compared to the UCB algorithm. In each round, the strategy chooses the arm with the largest value of A_t.

3.2. Novelty of Advanced MOSS Strategy

The novelty of this research is the introduction of an enhancement factor that can reduce the influence of uncertain arms as time wears on. The latter part of the new formula will decrease with the increase of time which means the strategy will focus more on the already known expected rewards information instead of unexplored knowledge. By doing so, this advanced algorithm can better balance the exploitation and exploration phases. As time goes on, the intensity of the exploration waned. This strategy can reduce the time cost of exploration without losing accuracy and increase the expected rewards because of less useless research. In the traditional MOSS algorithm, the strategy explores all the worst situations of each arm and relies too much on the exploration phase instead of the exploitation phase. It ignores the former information about each experiment round. So, this paper adds an enhancement factor ∂ in the denominator of the confidence bound formula to accelerate the exploration process. Here is the advanced MOSS algorithm formula:

$$A_t = arg\max_i \mu_i(t-1) + \sqrt{\frac{4}{T_i(t-1)^{\partial}} log^+\left(\frac{n}{kT_i(t-1)}\right)} \quad (2)$$

The factor ∂ is calculated as follows:

$$\partial = 1 + t/n \quad (3)$$

Where t is the experiment rounds that have been finished and n is the total experiment rounds. By adding such a factor, the confidence bound will become narrower and narrower as time goes on. Therefore, the strategy can focus more on the already known reward information and reduce the exploration time of arms that have large uncertainty. This prevents neglect to explore other potentially better arms. Besides, this can also stabilize the strategy performance and smooth the influence of noise reward points. So over time, the strategy will decrease the exploration intensity and better exploit the known information. Finally, the strategy realizes that it can fully utilize the former reward information and avoid unmeaning exploration in the meantime.

4. Experiment Design

For the experiment data set, this paper uses the MovieLens data set. It provides personalized movie recommendations and marks each movie. It begins with 100,000 ratings and then 25 million rating data has been expanded [14]. This research categorizes movies by their 18 kinds of genres and uses ratings as bandit arm rewards. As shown in figure 1, these 18 genres are chosen as arms and their average ratings are calculated as each arm's average rewards. Before the categorizing process, data preprocessing and data cleaning processes are essential. This process aims to remove the data that doesn't have attribute values, such as the movie_rating, movie_genre and so on. Then some error data strips are also removed:

DOI: 10.54254/2755-2721/83/2024GLG0076

because the rating attribute values are constrained from 0 to 5, the data that is not in this range is removed too.

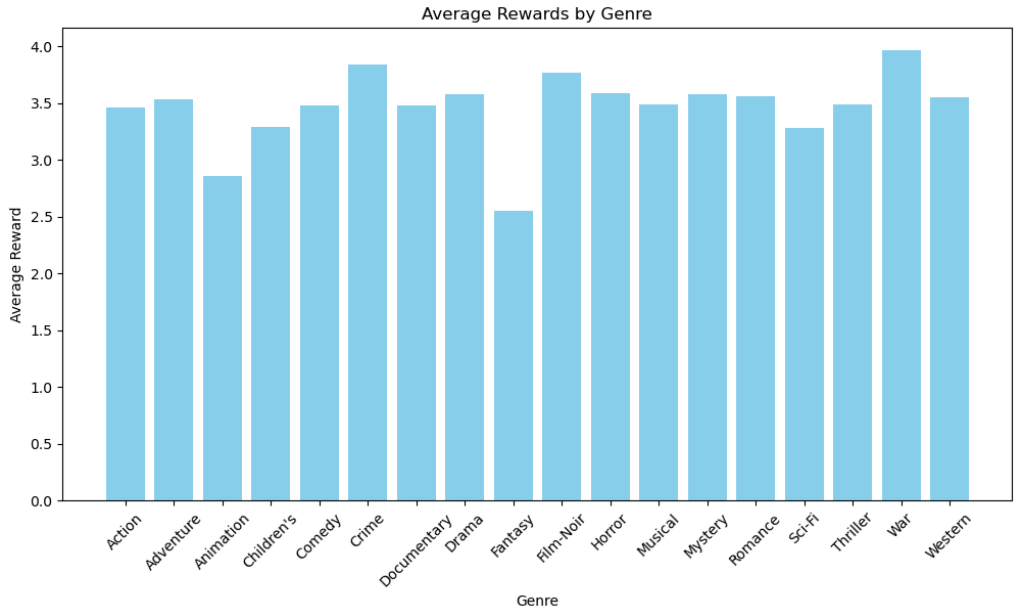

Figure 1. MovieLens dataset

After the preparatory work, this research applies normal UCB strategy, traditional MOSS strategy and advanced MOSS strategy. In each round, the decision-maker chooses one bandit arm according to the strategies and then gets a reward value (movie ratings). The strategies will continuously recommend movies based on the foregoing movie rating information and then update the known knowledge. Finally, the model will generate an evaluation index called cumulative regrets to compare the performance of algorithms. This index means the cumulative different values between the real best arm rewards and the selected arm rewards [15]. It can show the deviation between each strategy and the best strategy in each step. To avoid contingency, this paper conducts 10 experiments for each strategy and then averages these 10 cumulative regret values. This mean value is the final evaluation index, called the average cumulative regrets. The smaller the index, the better the algorithm. In every experiment, the paper sets the total round number n = 500,000.

For this data set, the advanced MOSS strategy should have better performance than the traditional MOSS strategy. It should have smaller average cumulative regrets and use fewer rounds to find the best arm because of its better exploration-exploitation balance.

5. Results

By comparing the average cumulative regrets between the advanced MOSS algorithm and other strategies, this experiment produces two figures to show the improvement of advanced MOSS. The first shows the improvement of the model accuracy which means lower average cumulative regrets. Another shows the fewer rounds to find the best arm. Although the UCB algorithm allows the optimism principle [11], its performance is even much worse than the normal MOSS strategy.

5.1. Result Figures

Figure 2 shows the relationship of average cumulative regrets over time of three strategies: UCB, traditional MOSS and advanced MOSS algorithm. Obviously, the UCB algorithm has a much worse performance than the other two algorithms. The key point is the comparison between two versions of MOSS strategies. As shown in Figure 2, the advanced MOSS strategy has smaller average cumulative regrets than the normal MOSS algorithm. This means that by adding the enhancement factor, the

DOI: 10.54254/2755-2721/83/2024GLG0076

advanced MOSS performs better than the traditional one and during the exploration-exploitation process, the advanced one has a larger chance to choose the best arm in each round than the original one. This proves that the advanced MOSS algorithm can truly improve the traditional algorithm and improve the model accuracy.

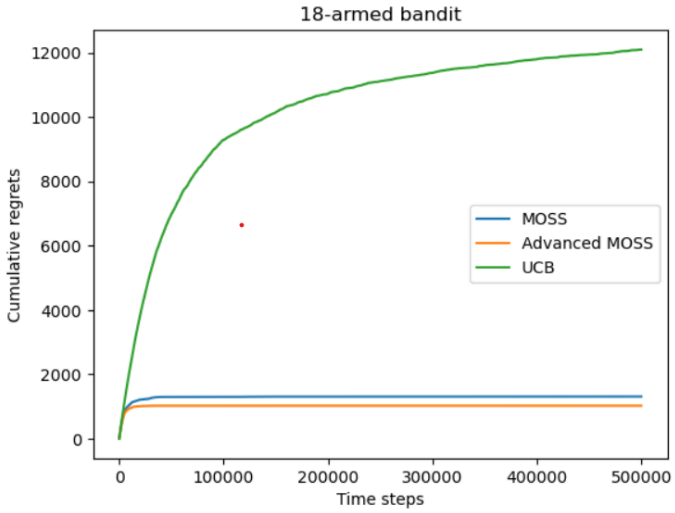

Figure 2. Average cumulative regrets (n=500,000)

Figure 3 shows the details of these two MOSS algorithms with total rounds n = 10,000. As shown in this figure, the advanced MOSS strategy not only has lower average cumulative regrets but also uses fewer rounds to reach the horizontal status. The advanced MOSS algorithm uses 2,000 time steps to find the best arm and gets a final regret which is near 300. But the traditional MOSS strategy uses almost 6,000 time steps to find the best arm and finally gets a regret of 550 which is almost twice that of the advanced MOSS regret. This proves that the advanced one can use less time to find the best strategy and find the most appropriate arm. It owes to the balance of exploration-exploitation phases. The introduction of the enhancement factor can reduce the probability of over-exploration and better exploit the former known knowledge.

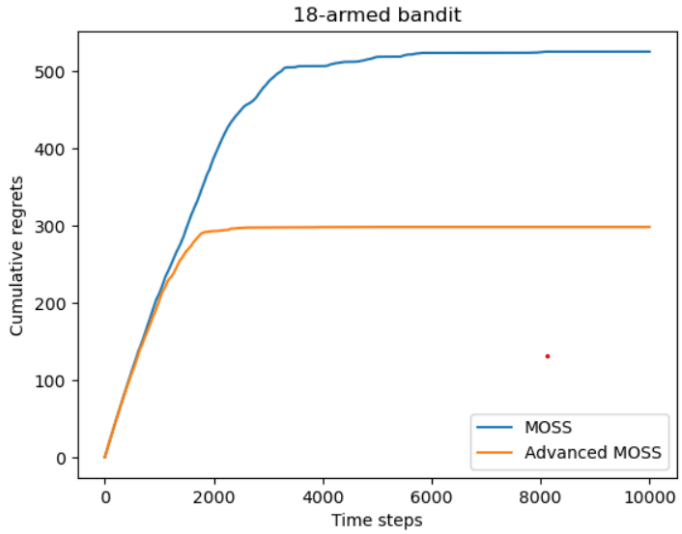

Figure 3. Comparison between MOSS and advanced MOSS

DOI: 10.54254/2755-2721/83/2024GLG0076

5.2. Result Comments

According to the results shown above, the advanced MOSS algorithm has about 800 cumulative average regrets which is much smaller than the 12,000 value of the UCB strategy (n = 500,000). Besides, with the passing of time steps, the MOSS algorithm tends to be horizontal, but the UCB algorithm will continue to rise slowly. The advanced MOSS algorithm has the smallest average cumulative regrets. Besides, it also uses fewer time steps to be horizontal which means it needs the least time to find the best arm and has the least error in the meantime. These all indicate that the advanced MOSS algorithm performs the best among all these strategies. This is reasonable because the advanced MOSS algorithm introduces an enhancement factor ∂ to balance the two model phases. Over time, as the factor becomes larger, the upper confidence bound becomes narrower and narrower. The model focuses more on the known reward information itself instead of trying to explore new arms with huge uncertainty. It allows this strategy to find the best arm with fewer rounds and higher accuracy. Besides, because of reducing unnecessary exploration steps, the cumulative expected regrets also become smaller which means this new advanced strategy can use less cost to achieve high efficiency. The advanced MOSS strategy can combine the advantages of exploration-exploitation phases and the benefits of minimax methods. That is why it can perform better than the traditional one and other bandit strategies.

6. Conclusion

This paper solves the over-exploration bandit problem by updating the advanced MOSS strategy and adding an enhancement factor. By introducing this factor, the strategy can reduce the cost of uncertain information in the later stages of the exploration phase. It will fully use the known expected rewards knowledge to make the optimal decisions without losing accuracy. This outcome is helpful for today's information age and can improve the recommended speed of Internet recommendation systems with less error. This research mainly focuses on the improvement of traditional MOSS algorithms. It mainly realizes one innovation point: add an enhancement factor ∂ into the confidence bound part. This advantage makes this advanced MOSS strategy perform better than before with fewer rounds needed. This new algorithm has much smaller average cumulative regrets and larger cumulative rewards than traditional MOSS and UCB strategies in the MovieLens data set. The result figures show that the advanced MOSS strategy can flatten out much faster than any other algorithms which means this strategy can use the least time to achieve the best results. The introduced enhancement factor plays an important role, and it can show the proportion of current rounds and total rounds. The larger this proportion, the less attention should be paid to new unknown information. This is workable because this new advanced strategy still leaves the model enough rounds to generate useful information and converge at a reasonable rate.

However, this strategy has two limitations. One is that it highly depends on the knowledge of the data set and the horizon of total rounds n. Further experiments like unsupervised machine learning methods [16] can be used to overcome it. Besides, a time step function should be used to replace the attribute of n. The other problem is that when calculating the new average reward of each arm in every round, the weight of the reward is equal. However, a real-time strategy should focus more on the new information. Because in the initial phase, the choice has great randomness. But with the passing of time, the arm chosen in each round is more likely to be the best one and so the information later should have a larger proportion than the former information. Further research will design a weighted average reward formula to update the whole data set information. By doing so, this strategy can overcome the problem of unknown time steps in practical applications. Moreover, the weighted formula can make the strategy prefer new knowledge instead of former erratic information. After these improvements, the MOSS strategy must be more efficient and realistic.

References

[1] Varatharajah, Y., & Berry, B. (2022). A contextual-bandit-based approach for informed decision-making in clinical trials. Life, 12(8), 1277.

DOI: 10.54254/2755-2721/83/2024GLG0076

[2] Elena, G., Milos, K., & Eugene, I. (2021). Survey of multiarmed bandit algorithms applied to recommendation systems. International Journal of Open Information Technologies, 9(4), 12-27.
[3] Foster, D. J., Kakade, S. M., Qian, J., & Rakhlin, A. (2021). The statistical complexity of interactive decision making. arXiv preprint arXiv:2112.13487.
[4] Krishnasamy, S., Sen, R., Johari, R., & Shakkottai, S. (2021). Learning unknown service rates in queues: A multiarmed bandit approach. Operations research, 69(1), 315-330.
[5] Zhu, Y., & Nowak, R. (2020). On regret with multiple best arms. Advances in Neural Information Processing Systems, 33, 9050-9060.
[6] Adusumilli, K. (2021). Risk and optimal policies in bandit experiments. arXiv preprint arXiv:2112.06363.
[7] Bertolini, M., Mezzogori, D., Neroni, M., & Zammori, F. (2021). Machine Learning for industrial applications: A comprehensive literature review. Expert Systems with Applications, 175, 114820.
[8] Xia, W., Quek, T. Q., Guo, K., Wen, W., Yang, H. H., & Zhu, H. (2020). Multi-armed bandit-based client scheduling for federated learning. IEEE Transactions on Wireless Communications, 19(11), 7108-7123.
[9] Wang, X., Jin, Y., Schmitt, S., & Olhofer, M. (2023). Recent advances in Bayesian optimization. ACM Computing Surveys, 55(13s), 1-36.
[10] Feng, J., Zhu, J., Zhao, X., & Ji, Z. (2024). Dynamic Grouping within Minimax Optimal Strategy for Stochastic Multi-ArmedBandits in Reinforcement Learning Recommendation. Applied Sciences, 14(8), 3441.
[11] Lu, Y., Xu, Z., & Tewari, A. (2021). Bandit algorithms for precision medicine. arXiv preprint arXiv:2108.04782.
[12] Feng, J., Zhu, J., Zhao, X., & Ji, Z. (2024). Dynamic Grouping within Minimax Optimal Strategy for Stochastic Multi-ArmedBandits in Reinforcement Learning Recommendation. Applied Sciences, 14(8), 3441.
[13] Lattimore, T., & Szepesvári, C. (2020). Bandit algorithms. Cambridge University Press.
[14] González, Á., Ortega, F., Pérez-López, D., & Alonso, S. (2022). Bias and unfairness of collaborative filtering based recommender systems in MovieLens dataset. IEEE Access, 10, 68429-68439.
[15] Lee, K., & Lim, S. (2022). Minimax optimal bandits for heavy tail rewards. IEEE Transactions on Neural Networks and Learning Systems, 35(4), 5280-5294.
[16] Wu, J., Braverman, V., & Yang, L. (2022, May). Gap-dependent unsupervised exploration for reinforcement learning. In International Conference on Artificial Intelligence and Statistics (pp. 4109-4131). PMLR.