Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/68/20241401

Optimizing video click-through rates with bandit algorithms

Ruhao Liu

School of Future Technology, South China University of Technology, Guangzhou, China

202264643089@mail.scut.edu.cn

Abstract. In recent years, videos have increasingly influenced public perception, making video platforms a focal point of digital consumption. One critical challenge for platform operators is identifying videos that resonate most with users, as user ratings directly reflect viewer preferences and experiences. This study explores the use of bandit algorithms to predict and strategize the overall ratings of various anime videos on the Bilibili platform. Bandit algorithms, a subset of the multi-armed bandit model, dynamically adjust selection strategies based on prior feedback to maximize cumulative rewards. Our empirical research assessed multiple gambling algorithms, including the ε-greedy, Upper Confidence Bound (UCB), Explore-then-Commit (ETC), and Thompson Sampling (TS) algorithms. The findings indicate that the Thompson Sampling algorithm, in particular, achieved the lowest cumulative regret in selecting optimal videos on the Bilibili platform, showcasing its superior performance. This study highlights the potential of bandit algorithms to enhance video selection processes, ensuring that platforms can effectively cater to user preferences and enhance viewer satisfaction.

Keywords: Thompson sampling, Reinforcement learning, multi-armed bandit, exploration and exploitation.

1. Introduction

In recent years, as videos increasingly influence people's lives, selecting user-preferred content has become a crucial decision-making challenge for platform operators. Intelligent algorithms are essential for constructing and implementing decisions in these domains. In machine learning, Reinforcement Learning demonstrates exemplary performance in known environments. Agents learn through actions, interacting with their surroundings, receiving feedback, and adjusting their strategies trial by trial to maximize future rewards. RL agents continually balance exploiting known information against exploring new data to optimize their strategy outcomes [1].

A classic problem formulation within RL is the Multi-Armed Bandit. MAB problems address sequential decision-making, with applications spanning clinical trials, finance, advertising, and recommendation systems [2]. The challenge involves making choices among multiple options, or "arms," each offering random rewards. By choosing an arm, the agent gains insights into that arm's characteristics and subsequently adjusts its strategy to balance exploitation of known rewards with exploration of other options. This strategy aims to maximize cumulative rewards by weighting the optimal arm more heavily [3]. The primary challenge lies in exploring optimal options while minimizing losses, thereby achieving a balance between exploitation and exploration in probabilistic

DOI: 10.54254/2755-2721/68/20241401

decision-making. This study utilizes data from Bilibili, a platform known for its anime videos, and employs regret as the metric to evaluate outcomes. We assess several classic bandit algorithms, including Upper Confidence Bound, Explore-then-Commit, and Thompson Sampling, alongside improved algorithms such as Adaptive UCB, to explore optimal algorithmic decisions in this environment. Our analysis aims to derive insightful conclusions based on the results, data factors, and the logic underpinning the algorithms.

2. Algorithm Analysis

2.1. Upper Confidence Bound (UCB)

The UCB algorithm is a classic algorithm in the context of the Multi-Armed Bandit (MAB) problem. During the process of action selection, it considers not only the historical average rewards of actions but also the confidence bounds of each action. The UCB algorithm selects the arm with the highest upper confidence bound based on the historical average rewards and confidence bounds of each arm.

$$\mathrm{UCB}_i(t-1) = u_i(t-1) + \frac{B}{2}\sqrt{\frac{4\log n}{T_i(t-1)}} \tag{1}$$

In the UCB algorithm, B represents the difference between the maximum and minimum possible reward values. In addition to the basic UCB algorithm, various derivative algorithms have been developed to address different scenarios. For example, Chen proposed the UCB-max algorithm, Roy referred to the UCB-KL algorithm and their introduction of the TV-KL-UCB algorithm, and Gil. proposed the UCB-RAD auxiliary algorithm, among others [4-6]. Considering the potential strong fluctuations in video rating data during execution, this paper also adopts the AUCB (Asymptotically Optimal UCB) algorithm to more accurately estimate the uncertainty of actions.

$$\mathrm{AUCB}_i(t-1) = u_i(t-1) + \frac{B}{2}\sqrt{\frac{2\log(f(t))}{T_i(t-1)}} \tag{2}$$

$$\text{Here } f(t) = 1 + t(\log t)^2. \tag{3}$$

2.2. Epsilon greedy

The Epsilon Greedy algorithm, compared to other MAB algorithms, is relatively straightforward. It relies on exponential weighting and does not require estimation of action uncertainty. Instead, it selects actions based on their historical performance and weights. The algorithm allocates a proportion of ϵ for exploration and $(1-\epsilon)$ for exploitation. As shown in Figure 1.

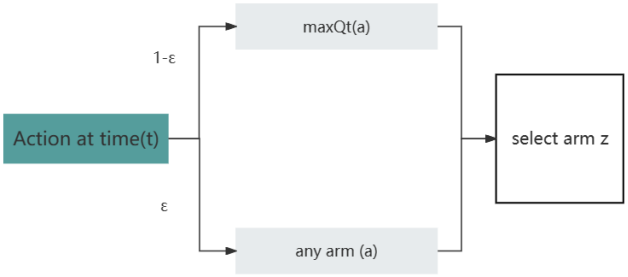

Figure 1. The Epsilon Greedy algorithm (Photo/Picture credit: Original).

The paper refers to a variant of the Epsilon Greedy algorithm proposed by Kuang, assuming that the presented extraction is an M-list, with each randomly selected arm denoted as Z-arm [7].

A. Re-selecting objects for exploration: If the given random object Z-arm is included in previous M-list demonstrations, it can be re-selected to be included in subsequent M-list demonstrations.

DOI: 10.54254/2755-2721/68/20241401

B. Not re-selecting objects for exploration: If the given random object Z-arm is included in previous M-list demonstrations, it will be excluded from subsequent M-list demonstrations [8].

Considering the relatively small Manhattan distance between video rating data, this paper selects EGSE-A. This variant exhibits greater robustness. Even if the agent ignores more optimal arms during the decision-making process, they can still be discovered in subsequent explorations. However, this variant often slows down the exploration process.

Additionally, this paper introduces the Decaying Epsilon Greedy algorithm, which incorporates a decay factor as a new variable. The decay factor is defined as DF.

$$\varepsilon(\mathrm{t}-1)=\frac{1}{\mathrm{T}(\mathrm{t}-1)^{\mathrm{DF}}} \tag{4}$$

Compared to the Epsilon Greedy algorithm, the Decaying Epsilon Greedy algorithm can dynamically adjust decisions while increasing the algorithm's convergence speed, thus reducing losses due to exploration.

2.3. Thompson Sampling (TS)

Thompson Sampling is a Bayesian-inference-based method that performs random sampling according to the posterior distribution. Traditional Thompson Sampling typically uses the Beta distribution as the probability distribution for rewards. The Beta distribution takes the number of successes and failures as parameters (α and β). As one of these parameters increases, the peak of the distribution shifts towards 1 or 0.

However, in this experiment, the data's contrast is poor. Gaussian distribution is used as the reward distribution instead. Here, B represents the difference between the maximum and minimum possible rewards.

$$F_i(t)\ \sim N(u_i(t),\frac{B^2}{4T_i(t)}) \tag{5}$$

Using Gaussian distribution for reward distribution reflects the continuous nature of rating data more effectively, thus highlighting the characteristics of reward differences. Consequently, the intelligent agent can better discern the optimal arm with greater accuracy.

2.4. Explore-Then-Commit (ETC)

The ETC algorithm is a more intuitive approach to addressing the exploration-exploitation trade-off in the MAB problem. The algorithm is intuitively divided into two stages: exploration and exploitation. During the exploration stage, the algorithm directly selects arms that have not been observed before to explore. After the exploration stage, the optimal arm is identified, and during the exploitation stage, this arm is executed. The algorithm manages the balance between exploration and exploitation by using parameters k to represent the number of arms explored and m to represent the depth. This approach directly addresses the balance between exploration and exploitation.

3. Experimental Approach

3.1. Selection of Elements and Procedures

In this experiment, web scraping techniques were utilized to collect comprehensive rating data of anime videos from the BiliBili website. The dataset comprises ratings for nearly four thousand anime videos. Ratings are on a scale of ten, and after cleaning out irrelevant data, a total of 28 video types remained. Fig 2 illustrates the distribution of average ratings and data volume in this dataset.

DOI: 10.54254/2755-2721/68/20241401

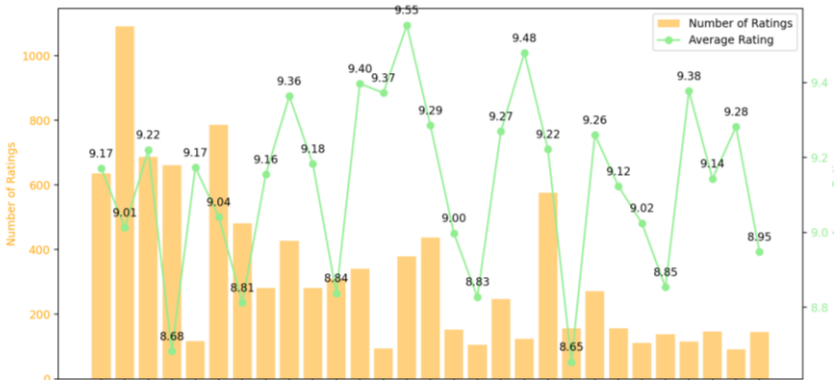

Figure 2. The distribution of average ratings and data volume (Photo/Picture credit: Original).

Prior to experiment commencement, the average reward for the highest-rated type among the best arm rating data was set as the optimal expected value. Regret, on the other hand, represents the difference between the optimal expected value and the reward value returned each time. Each algorithm in this study conducted a total of 100,000 extractions, repeated 10 times. Throughout this process, we recorded the reward value for each extraction and concurrently calculated the mean and variance to plot error bars, visually reflecting the accuracy of the algorithm in determining the optimal type and its stability.

3.2. Prediction of Results

Including improved algorithms, this experiment employed six different algorithms. Based on Mambou's understanding of the Epsilon algorithm, we set it accordingly [9]. B was set to 4.4 based on the distribution of the data. Since the data itself does not dynamically update according to the experiment, the ETC algorithm is expected to be more stable compared to other algorithms. AUCB algorithm, not relying on prior assumptions about reward distributions, is more suitable for a wider range of reward distributions, such as the ten-point rating scale data [10]. The TS algorithm samples arms based on posterior distributions, making it flexible and effective in most scenarios. Therefore, it is anticipated in this study that AUCB, ETC, and TS algorithms will demonstrate more effective performance in datasets with broader reward distributions.

4. Results Analysis

After parameter optimization and some adjustments, this study obtained the results depicted in Figure 3. During the algorithm adjustment process, it was observed that some algorithms only demonstrated good convergence when a sufficient number of samples were collected. UCB and AUCB algorithms showed noticeable convergence trends at n = 35000. TS algorithm exhibited convergence at n = 1000, while Epsilon greedy algorithm started to show relatively objective convergence trends at n = 150000

In the setup of the Decaying Epsilon greedy algorithm, particular attention needs to be paid to the impact of the decay factor. Adjusting the percentile can also have a significant impact on the algorithm's results. From the results, the TS algorithm demonstrates the best performance. It not only converges quickly but also exhibits the lowest cumulative regret among the six algorithms. This is consistent with the predicted outcome, as the TS algorithm updates posterior distributions using Bayesian inference. By utilizing Gaussian distributions for reward distributions, the TS algorithm is better able to capture the overall shape and nuanced differences in reward value distributions across different types.

While the UCB algorithm exhibits stability and demonstrates effective convergence, it requires generating a significant amount of loss before converging. This suggests that the UCB algorithm is suitable for scenarios requiring extensive data for training and long-term decision-making contexts.

DOI: 10.54254/2755-2721/68/20241401

Although the AUCB algorithm exhibits differences compared to the predicted results, it still demonstrates its applicability when facing video rating data with a wider distribution, especially when compared to the UCB algorithm.

The Epsilon greedy algorithm and the Decaying epsilon greedy algorithm with a decay factor both demonstrate good performance. The regret values of both algorithms are consistently below 5000 across ten experiments. However, the Epsilon greedy algorithm fails to exhibit convergence. This is attributed to the fact that regardless of whether the optimal arm is correctly identified, the algorithm always has a fixed probability of ten percent to explore other arms. This results in the algorithm wasting considerable time exploring actions that are not accurate enough. While the Decaying epsilon greedy algorithm shows promising performance in some cases, its large error bars indicate poor stability. The viewpoint of this paper is that the algorithm's decay rate is too fast, leaving insufficient space for exploration, thereby reducing the accuracy of selecting the optimal arm.

The ETC algorithm demonstrates strong stability when its outcomes align with predictions. Although there is some loss during the exploration phase, its ability to identify the best arm is excellent. This stability is particularly notable despite the inevitable losses incurred during exploration.

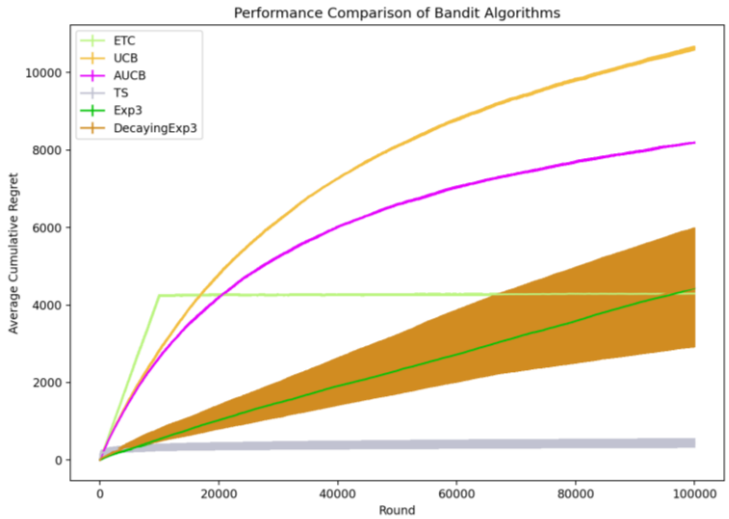

Figure 3. Performance Comparison of Bandit Algorithms (Photo/Picture credit: Original).

5. Conclusion

This study undertook an empirical evaluation using rating data from various anime videos on the Bilibili platform to assess the feasibility and effectiveness of using gambling algorithms for comprehensive rating predictions and strategy formulation. We analyzed several traditional algorithms—such as the ε-greedy, Upper Confidence Bound (UCB), Explore-then-Commit (ETC), and Thompson Sampling (TS) algorithms—as well as advanced variants like the Adaptive UCB (AUCB) and Decaying ε-greedy algorithms. Our findings revealed that the Thompson Sampling algorithm consistently achieved the lowest cumulative regret when selecting optimal elements, underpinned by its use of Bayesian updates and Gaussian reward distributions to refine the accuracy of reward estimations across different video types. This facilitated more effective decision-making. Conversely, the UCB algorithm, despite its stable convergence, required extensive data and incurred notable losses. The AUCB algorithm, while useful for datasets with broader rating distributions, still showed limitations compared to the traditional UCB. Both the ε-greedy and Decaying ε-greedy algorithms performed well in our tests; however, the ε-greedy algorithm suffered from poor convergence, potentially leading to prolonged engagement with suboptimal choices. The Decaying ε-greedy algorithm, affected by its rapid decay rate, struggled with stability, impacting its accuracy in identifying the best arm. The ETC algorithm displayed robust performance and was particularly adept at pinpointing the optimal choice, despite some losses during

DOI: 10.54254/2755-2721/68/20241401

the exploration phase. Looking ahead, given the dynamic nature of user ratings in real-world settings, we plan to continuously update our dataset and focus on enhancing the TS and UCB algorithms. Our goal is to develop more adaptable and responsive algorithms capable of real-time prediction and adjustment of video ratings.

References

[1] Mambou, E. N., & Woungang, I. (2023). Bandit Algorithms Applied in Online Advertisement to Evaluate Click-Through Rates. 2023 IEEE AFRICON, Nairobi, Kenya, 1-5.

[2] Bouneffouf, D., & Rish, I. (2019). A survey on practical applications of multi-armed and contextual bandits. arXiv preprint. Retrieved from https://arxiv.org/abs/1909.12335

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

[4] Cesa-Bianchi, N. (2006). Prediction learning and games. Cambridge University Press.

[5] Kao, K.-Y., & Chen, I.-H. (2014). Maximal expectation as upper confidence bound for multi-armed bandit problems. IEEE 7th Joint International Information Technology and Artificial Intelligence Conference, Chongqing, China.

[6] Roy, S., Shakkottai, S., & Srikant, R. (2024). Adaptive KL-UCB Based Bandit Algorithms for Markovian and I.I.D. Settings. IEEE Transactions on Automatic Control, 69(4), 2637-2644.

[7] Auer, P. (2012). Using confidence bounds for exploitation-exploration trade-offs. Journal of Machine Learning Research, 3(Nov), 397-422.

[8] Gil, Y., Baek, J., Park, J., & Han, S. (2021). Automatic Data Augmentation by Upper Confidence Bounds for Deep Reinforcement Learning. 2021 21st International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea, Republic of, 1199-1203.

[9] Kuang, N. L., & Leung, C. H. C. (2019). Performance Effectiveness of Multimedia Information Search Using the Epsilon-Greedy Algorithm. 2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA), Boca Raton, FL, USA, 929-936.

[10] Li, M. (2021). Search of Multiple Targets with Different Unknown Distributions Using Thompson Sampling. 2021 3rd International Symposium on Robotics & Intelligent Manufacturing Technology (ISRIMT), Changzhou, China, 207-212.