Advancing decision-making strategies through a comprehensive study of Multi-Armed Bandit algorithms and applications

Yang Kuang

DUT School of Software Technology & DUT-RU International School of Information Science and Engineering, Dalian University of Technology, Dalian, China

13502287010@mail.dlut.edu.cn

Abstract. Multi-Armed Bandit (MAB) strategies play a pivotal role in decision-making algorithms by adeptly managing the exploration-exploitation trade-off in environments characterized by multiple options and constrained resources. This paper delves into the core MAB algorithms, including Explore-Then-Commit (ETC), Thompson Sampling, and Upper Confidence Bound (UCB). It provides a detailed examination of their theoretical underpinnings and their application across diverse sectors such as recommender systems, healthcare, and finance. MAB algorithms are celebrated for their efficiency in optimizing decision outcomes; however, they are not without challenges. Significant issues include managing the complexity of exploration and adapting to non-stationary environments where the dynamics of the available options may change over time. A nuanced understanding of these challenges is crucial for effectively implementing MAB strategies in complex decision-making scenarios. This study not only highlights the versatility and potential of MAB algorithms but also underscores the need for ongoing research to refine these techniques and expand their applicability.

Keywords: Multi-Armed Bandit, Explore-Then-Commit, Thompson Sampling, Upper Confidence Bound.

1. Introduction

Multi-Armed Bandit (MAB) strategies have become essential in the evolving landscape of decision-making algorithms, addressing the crucial challenge of balancing exploration with exploitation where resources are limited and must be allocated among multiple options. These strategies are central to optimizing outcomes across various sectors, including recommender systems, healthcare, and finance.

The essence of MAB algorithms is the exploration-exploitation dilemma [1], akin to a gambler who must choose between continuing with a known profitable slot machine (exploitation) or trying out others in search of potentially greater rewards (exploration). MAB algorithms dynamically adjust this balance by learning from past interactions to maximize cumulative rewards. Key strategies in MAB research include the Explore-Then-Commit (ETC) approach, Thompson Sampling, and Upper Confidence Bound (UCB) [2]. The ETC strategy involves an initial exploration of various options, subsequently committing to the one that has demonstrated the best performance based on gathered data [3]. Thompson Sampling employs Bayesian inference to dynamically adjust selections based on probabilistic estimations of each option's potential [4], while UCB focuses on options with high uncertainty to maximize expected gains. The effectiveness of these strategies is often measured by regret, defined as the opportunity cost incurred by not choosing the optimal option. Establishing regret bounds allows for a quantitative evaluation of algorithm performance.

In practical applications, MAB algorithms enhance personalization in e-commerce, streaming services, and news delivery by adapting recommendations in real time to user preferences [5]. In healthcare, they are utilized for optimizing clinical trials, drug dosages, and treatment plans, balancing well-known treatments with experimental options [6]. In finance, MAB strategies support portfolio management, dynamic pricing, and advertising bidding to optimize returns while managing risks [7]. Despite their efficacy, MAB algorithms encounter challenges such as managing exploration complexity, adapting to non-stationary environments, and improving sample efficiency [8]. This study bridges the gap between theoretical understanding and practical implementation, providing decision-makers with the tools to navigate the complexities of the MAB landscape effectively. In conclusion, Multi-Armed Bandit strategies are highly adaptable and effective, proving indispensable across multiple domains. They continue to drive advancements in decision-making and resource optimization. As research progresses, ongoing innovations are expected to address current challenges and expand the scope of MAB applications.

2. Relevant Theories

2.1. Basic Principles of Multi-Armed Bandit Algorithms

$$R_T = T\mu^* - \sum_{t=1}^{T} \mu_{j(t)} \tag{1}$$

Where $\mu^* = max_{i=1,2,3\cdots k}\mu_i$ is the expected reward from the best arm.

Alternatively, the total expected regret can be expressed as

$$R_T = T\mu^* - u_{j(t)} \sum_{k=1}^{K} E\big(T_k(T)\big)$$

$$\mathrm{E}(\mathrm{T_k}(\mathrm{T})) \geq \frac{\ln \mathrm{T}}{\mathrm{D}(\mathrm{p_k}||\mathrm{p^*})} \tag{2}$$

$$\mathrm{D}(\mathrm{p_k}||\mathrm{p^*}) = \int \mathrm{p_j} \ln \frac{\mathrm{p_j}}{\mathrm{p^*}}$$

Regret grows at least logarithmically, or more formally, $R_T = \Omega(\log T)$. An algorithm is considered to have solved the multi-armed bandit problem if it can meet the lower bound of $R_T = O(\log T)$.

2.2. Exploration vs. Exploitation Trade-off

In decision problems, exploitation and exploration are two possible behaviours, each with its advantages and disadvantages. On the one hand, exploitation refers to taking what is considered to be the optimal decision based on the data observed so far. This 'safe' approach avoids bad decisions as much as possible, but also prevents the possibility of discovering potentially better decisions. Exploration, on the other hand, involves not taking what appears to be the optimal decision, but rather betting that the observed data is insufficient to truly identify the best option. This more' risky' approach can sometimes lead to poor decisions, but it can also lead to the discovery of better options if they exist.

The choice development versus exploration problem arises in many cases where observations lead to decisions and decisions lead to new observations. Faced with such "feedback loops" (data generate decisions, decisions generate data), we must always ask ourselves whether the apparently optimal decision is really optimal, or whether the observed data are not representative enough to identify the truly optimal decision. The goal is then to find a strategy that is a good compromise between these two behavioural approaches [9].

A unique challenge in reinforcement learning is to balance the relationship between exploring the environment and using the agent's knowledge to maximise the cumulative reward. Typically, to maximise cumulative reward, agents should prioritise actions that have previously produced positive results. However, in order to discover new effective actions, agents must also explore and try less discovered options. Neither exploration nor exploitation will be successful if the task is focused on exploration alone. In the case of pure exploration, the agent will always choose actions that provide more information about the environment. Gradually, the agent will learn more about the environment and which actions to take.

2.3. Classical MAB Algorithms

The Explore-Then-Commit (ETC) algorithm is a strategy for solving Dobby slot machine problems. Then, in the commitment phase, the algorithm chooses the action that performed best in the exploration phase and continues to choose that action for the rest of the time.

The key to the ETC algorithm is the balance between exploration and exploitation. If the exploration time is too short, it may not be able to gather enough information to make the best decision; if the exploration time is too long, it may miss the opportunity to exploit the best action for a higher reward. Let $\widehat{\mu_\iota}(t)$ be the average reward received from arm i after round t, which is written formally as.

$$\widehat{\mu_\iota}(t) = \frac{1}{T_i(t)} \sum_{s=1}^{t} \mathrm{II}\{A_s = i\}\ X_S \tag{3}$$

1: Input m $\in$N.

2: In round t choose action

$$A_t = \begin{cases} i, & if\ (t\ modK) + 1\ =\ i\ and\ t\ \leq\ mK; \\ & argmax_i \widehat{\mu_\iota}(t)\ (mK), t\ > mK. \end{cases} \tag{4}$$

The expected regret of the ETC policy is bounded by,

$$R_T \leq m \sum_{i=1}^{k} \Delta_i + (n - mk) \sum_{i=1}^{k} \Delta_i exp(-\frac{m\Delta_i^2}{4}) \tag{5}$$

Initially, the algorithm has no knowledge of the rewards of the different arms. So it starts by selecting each arm once to get an initial estimate of the rewards. It ensures that arms that have been selected less frequently (and therefore have more uncertainty) are given a chance.

After an arm has been selected and rewarded, the algorithm updates the empirical mean reward for that arm, repeating steps 2 and 3 for the desired number of trials.

The exploration term in the UCB algorithm is crucial because it balances the need to exploit arms that have historically provided high rewards (exploitation) with the need to explore arms that have not been selected as often (exploration). This balance is essential for dealing with the exploration-exploitation trade-off, which is a key challenge in reinforcement learning.

The selection strategy of the UCB algorithm is as follows: in each round, choose the action that maximises the following formula

$$\mathrm{UCB} = \mathrm{X} + \sqrt{\frac{2 \ln \mathrm{N}}{\mathrm{n}}} \tag{6}$$

The first term X on the right side of this formula represents exploitation, and the second term $\sqrt{\frac{2 \ln \mathrm{N}}{\mathrm{n}}}$ represents exploration. When an action is tried less frequently, the second term is larger, encouraging the algorithm to try actions that have been tried less often.

To evaluate the performance of the UCB algorithm, the most important thing is to analyze its regret. Regret is calculated using the following formula:

$$\mathrm{Regret} = \sum_{\mathrm{j}:\mu_\mathrm{j}<\mu^*} \Delta_\mathrm{j} \mathrm{E}[\mathrm{T_j}(\mathrm{n})] \tag{7}$$

One important property of the UCB algorithm is that its regret grows at a logarithmic rate, which means that the long-term performance of the UCB algorithm is very good. Specifically, for the original UCB algorithm, the upper bound of its regret after T rounds is $O(K \log T)$, where K is the number of actions. This means that although we cannot always make the optimal choice, by using the UCB algorithm, we can ensure that our regret grows very slowly, thereby obtaining a higher total reward in the long term.

Thompson Sampling is a reinforcement learning algorithm based on Bayesian thinking, mainly used to solve the multi-armed bandit problem. The basic idea is to model the probability of each action's reward, and then at each step, sample according to the current probability distribution and execute the action with the maximum sample value.

Specifically, suppose we have a context environment $x \in X$, make an action $a \in A$, and get a reward $r \in R$, then the likelihood function of this reward is $p(r|\theta, \alpha, x)$, where $\theta \in \Theta$ is the parameter of the reward distribution.

Assuming we know the prior probability distribution $p(\theta)$, suppose we have historical observation triplets D={(x,a,r)}, so the posterior distribution can be calculated $p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$.

The expected reward $E(r(\theta, \alpha, x)) = \int_{\theta} I[E(r|\theta, \alpha, x) = \max_{\alpha'} E(r|\theta, \alpha', x) p(\theta|D) d\theta$.

The reason why it is called sampling is because in actual operation, at each round of iteration, we get a $p(\theta|D)$, get a θ^* through the way of sampling, and then get $a^* = argemax_a E[r|\theta^*, a, x]$.

In Bayesian probability theory, $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(x|\theta')p(\theta')d\theta'}$,where θ is the model parameter, x is the model observation value, also known as evidence, $p(\theta|x)$ is called the posterior distribution, $p(x|\theta)$ is called the likelihood function, $p(\theta)$ is called the prior distribution, p(x) is called the marginal likelihood function.

3. System Analysis and Application Research

3.1. Recommender Systems

The combination of recommender systems and bandit algorithms shows potential for enhancing recommendation effectiveness. Traditional hybrid systems have mainly focused on combining recommendation algorithms as black boxes. However, recent attention has shifted towards dynamic ensemble improvement using bandit techniques. Bandit recommender system ensembles use multi-armed bandit strategies to select the next item for recommendation based on past user interactions. This approach has several advantages, including adaptability to changing user preferences and efficient resource allocation.

Recommender system ensembles are typically formed by combining individual recommendation algorithms to improve overall performance. However, this approach often requires extensive parameter tuning and computational resources, which limits its scalability in real-world scenarios. Bandit recommender system ensembles tackle these challenges by dynamically selecting the most effective recommendation algorithm for each user interaction, optimizing recommendation quality while minimizing computational overhead.

Algorithms are updated based on received rewards, either after each recommendation or in batches. The selection of the algorithm for each recommendation is based on its historical performance. Commonly employed strategies include ε-greedy and Thompson sampling.

The ε-greedy algorithm balances exploration and exploitation by selecting the algorithm with the highest average reward most of the time, while occasionally exploring other algorithms with a probability ε. In contrast, Thompson sampling models the uncertainty of each algorithm's reward distribution using a Beta distribution and selects the algorithm with the highest sampled value.

Bandit recommender system ensembles offer several advantages over traditional ensemble approaches. The system adapts in real-time to user feedback, ensuring that recommendations remain relevant and effective. Moreover, it is well-suited for A/B testing, automatically allocating traffic to different algorithms based on their performance. Unlike traditional A/B testing, where traffic to underperforming algorithms is gradually reduced, bandit ensembles allow for continuous adaptation, ensuring optimal performance even in non-stationary environments.

In summary, bandit recommender system ensembles are a new approach to improving recommendation quality while maintaining scalability and adaptability. These ensembles use multi-armed bandit strategies to dynamically select the most effective recommendation algorithms for each user interaction, leading to enhanced user satisfaction and engagement.

3.2. Healthcare Applications

Clinical trials are essential for testing the efficacy of various treatments. However, researchers often face the challenge of resource allocation, which involves judiciously distributing limited resources such as patients, funding, and time among competing treatment arms to maximize benefits while minimizing risks. To address this challenge, Multi-Armed Bandit (MAB) algorithms provide an adaptive solution. The algorithms used in this study treat each treatment arm as a potential lever and dynamically allocate resources based on observed outcomes.

For example, employed contextual bandit algorithms to optimize clinical trial design [10]. They efficiently compared treatment options by leveraging Gaussian Process regression and sub-sampling techniques. The use of this adaptive allocation strategy facilitated more effective data collection, enabling informed decision-making throughout the trial process.

In the realm of personalised medicine, determining optimal drug dosages for individual patients poses a significant challenge due to patient variability in characteristics and genetics. Multi-armed bandit (MAB) algorithms offer a promising solution by exploring different dosages while exploiting known information, thus enhancing patient outcomes.

Using a multi-armed bandit framework, addressed the dosage optimization problem of warfarin, a commonly used oral anticoagulant [11]. Their algorithm, which is based on the LASSO estimator, assigns appropriate initial dosages to patients efficiently. This personalized approach minimizes the risks associated with incorrect initial dosages, thereby improving patient safety and treatment efficacy.

Understanding reward processing abnormalities in various mental disorders, such as Parkinson’s, Alzheimer’s, ADHD, addiction, and chronic pain, is crucial for effective treatment [12]. MAB algorithms offer a dynamic approach to explore and exploit different treatment strategies tailored to individual patient needs. It is important to note that this approach can be particularly effective in treating patients with these conditions.

In conclusion, Multi-Armed Bandit algorithms provide versatile solutions in various healthcare domains. They optimize clinical trial design, personalize drug dosages, and address reward processing abnormalities in mental disorders. By dynamically balancing exploration and exploitation, these algorithms enable researchers and clinicians to make more informed and adaptive decisions, ultimately enhancing patient care and treatment outcomes.

3.3. Financial Sector Applications

Multi-Armed Bandit algorithms are useful in financial decision-making, particularly in portfolio optimization and dynamic pricing strategies.

Portfolio management involves allocating capital across various assets, such as stocks and bonds, to maximize returns while mitigating risks. MAB algorithms provide a dynamic framework for making online portfolio choices. These algorithms use correlations among multiple assets to adjust investments based on observed outcomes. The aim is to balance passive investments, such as index funds, with active strategies like stock picking, while considering risk-adjusted rewards.

For example, proposed a bandit algorithm to construct orthogonal portfolios from diverse assets [13]. By integrating the approach with the upper-confidence-bound bandit framework, an optimal portfolio strategy was derived. In addition, incorporated risk-awareness into the classic MAB setting by filtering assets based on the financial market's topological structure and combining the optimal bandit policy with coherent risk measures [14].

In the domain of dynamic pricing, online retailers face the challenge of setting real-time prices for their products to maximize long-term profits. MAB algorithms provide a solution for guiding dynamic pricing decisions. They allow companies to conduct price experiments that balance immediate profit with future profitability.

Proposed a dynamic price experimentation policy that incorporates MAB techniques and partial identification of consumer demand from economic theory, even with incomplete demand information [15]. Addressed dynamic pricing for multiple products by reducing the revenue maximization problem to an online bandit convex optimization [16]. Their approach incorporates latent product features and online singular value decomposition, leveraging side information from observed demands.

MAB algorithms are invaluable tools for optimizing financial decisions in portfolio management and pricing strategies. These algorithms enable investors and businesses to make data-driven decisions, resulting in more efficient allocation of resources and maximization of profits.

4. Challenges

Although Multi-Armed Bandit algorithms offer promising solutions across various domains, they still face several challenges that require further research and development.

One of the main challenges is balancing exploration and exploitation effectively, especially in complex environments with numerous options and uncertain outcomes. Therefore, developing efficient exploration strategies that minimize regret while maximizing long-term rewards remains an active area of research.

Sample Efficiency: MAB algorithms typically require a large number of samples to accurately estimate the rewards of each arm, especially in high-dimensional or sparse reward spaces. Improving sample efficiency to reduce the computational burden and expedite learning is crucial for scaling MAB techniques to real-world applications.

Additionally, MAB algorithms depend on specific assumptions about the underlying reward distributions, such as independence and identically distributed (i. i.d.) samples. Ensuring the robustness of MAB techniques to violations of assumptions and developing adaptable algorithms for diverse reward structures are critical for practical deployment.

Scalability is a significant concern as MAB algorithms are applied to larger datasets and more complex decision-making problems. Efficient algorithm design and optimization techniques are crucial for practical deployment, especially when dealing with large-scale problems.

Interdisciplinary research efforts combining insights from machine learning, optimization, statistics, and domain-specific knowledge are necessary to address these challenges. MAB algorithms have the potential to offer innovative solutions to a wide range of decision-making problems, provided that the obstacles are overcome.

5. Conclusion

Multi-Armed Bandit strategies are essential for addressing the exploration-exploitation trade-off in decision-making algorithms across diverse domains. MAB algorithms dynamically balance the need to explore new options and exploit known ones, optimizing outcomes in scenarios with multiple choices and limited resources. This paper presents an extensive review of basic MAB algorithms, such as Explore-Then-Commit, Thompson Sampling, and Upper Confidence Bound, discussing their theoretical foundations and practical applications in recommender systems, healthcare, and the financial sector. Although MAB algorithms offer significant benefits, they still face challenges such as exploration complexity, non-stationary environments, and sample efficiency. However, ongoing research and innovation are expected to address these challenges and uncover new applications for Multi-Armed Bandit techniques. In summary, MAB strategies represent a powerful approach to decision-making and resource optimization, driving advancements across various fields. As our understanding of MAB algorithms continues to evolve, they are poised to play an increasingly vital role in shaping the future of intelligent systems and data-driven decision-making.

References

[1] Galichet, N., Sebag, M., & Teytaud, O. (2013, October). Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In Asian conference on machine learning (pp. 245-260). PMLR.

[2] Garivier, A., Lattimore, T., & Kaufmann, E. (2016). On explore-then-commit strategies. Advances in Neural Information Processing Systems, 29.

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

[4] Garivier, A., & Moulines, E. (2011, October). On upper-confidence bound policies for switching bandit problems. In International conference on algorithmic learning theory (pp. 174-188). Berlin, Heidelberg: Springer Berlin Heidelberg.

[5] Silva, N., Werneck, H., Silva, T., Pereira, A. C., & Rocha, L. (2022). Multi-armed bandits in recommendation systems: A survey of the state-of-the-art and future directions. Expert Systems with Applications, 197, 116669.

[6] Zhou, T., Wang, Y., Yan, L., & Tan, Y. (2023). Spoiled for choice? personalized recommendation for healthcare decisions: A multiarmed bandit approach. Information Systems Research, 34(4), 1493-1512.

[7] Zhu, X., Huang, Y., Wang, X., & Wang, R. (2023). Emotion recognition based on brain-like multimodal hierarchical perception. Multimedia Tools and Applications, 1-19.

[8] Kuleshov, V., & Precup, D. (2014). Algorithms for multi-armed bandit problems. arXiv preprint arXiv:1402.6028.

[9] Le, D. H. (2023). Exploration-Exploitation Trade-off Approaches in Multi-Armed Bandit.

[10] Durand, A., Achilleos, C., Iacovides, D., Strati, K., Mitsis, G. D., & Pineau, J. (2018, November). Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In Machine learning for healthcare conference (pp. 67-82). PMLR.

[11] Bastani, H., & Bayati, M. (2020). Online decision making with high-dimensional covariates. Operations Research, 68(1), 276-294.

[12] Bouneffouf, D., & Rish, I. (2019). A survey on practical applications of multi-armed and contextual bandits. arXiv preprint arXiv:1904.10040.

[13] Shen, W., Wang, J., Jiang, Y. G., & Zha, H. (2015, June). Portfolio choices with orthogonal bandit learning. In Twenty-fourth international joint conference on artificial intelligence.

[14] Huo, X., & Fu, F. (2017). Risk-aware multi-armed bandit problem with application to portfolio selection. Royal Society open science, 4(11), 171377.

[15] Misra, K., Schwartz, E. M., & Abernethy, J. (2019). Dynamic online pricing with incomplete information using multiarmed bandit experiments. Marketing Science, 38(2), 226-252.

[16] Mueller, J. W., Syrgkanis, V., & Taddy, M. (2019). Low-rank bandit methods for high-dimensional dynamic pricing. Advances in Neural Information Processing Systems, 32.