Transactions on Computer Science and Intelligent Systems Research
ISSN: 2960-1800, eISSN: 2960-2238 | Volume 5
AIDML 2024

Real-world Applications of Bandit Algorithms: Insights and Innovations

Qianqian Zhang

School of Remote Sensing Information Engineerin, Wuhan University, Wuhan, 430072, China

2021302131219@whu.edu.cn

Abstract. In the rapidly evolving landscape of decision-making systems, the significance of Multi-Armed Bandit (MAB) algorithms has surged, showcasing a remarkable ability to address the exploration-exploitation dilemma across diverse domains. Originating from the probabilistic and statistical decision-making framework, MAB algorithms have established a critical role by offering a systematic approach to making choices in uncertain environments with limited information. These algorithms ingeniously balance the trade-off between exploiting known resources for immediate gains and exploring new possibilities for future benefits. The spectrum of MAB algorithms ranges from Stochastic Stationary Bandits, dealing with static reward distributions, to more complex forms like Restless and Contextual Bandits, each tailored to the dynamism and specificity of real-world challenges. Further, Structured Bandits explore the underlying patterns in reward distributions, providing strategic insights into decision-making processes. The practical applications of these algorithms span several fields, including healthcare, content recommendation, and education, demonstrating their versatility and efficacy in addressing specific contextual challenges. This paper aims to provide a comprehensive overview of the development, nuances, and practical applications of MAB algorithms, highlighting their pivotal role in advancing decision-making processes amidst uncertainty.

Keywords: Multi-armed bandit, real-world applications, recommendation system.

1. Introduction

The advent of Multi-Armed Bandit (MAB) algorithms marks a significant milestone in the evolution of decision-making frameworks, especially in environments characterized by uncertainty and incomplete information. At the core of these algorithms is the exploration-exploitation dilemma—a fundamental challenge that requires balancing the use of existing knowledge to maximize immediate rewards against taking exploratory actions that may yield future benefits. This conundrum is emblematic of numerous real-world situations where decisions must be made under the shadow of uncertainty, rendering MAB algorithms an invaluable tool across a broad spectrum of applications.

The evolution of MAB algorithms is characterized by their diversification to address complex decision-making scenarios. Stochastic Stationary Bandits focus on environments with stable reward probabilities, necessitating a balance between exploiting known options and exploring new ones. Meanwhile, Restless Bandits adapt to environments where reward probabilities evolve over time, demanding continuous strategic adjustments. Each variant of the MAB problem introduces new dimensions to the exploration-exploitation trade-off. Contextual Bandits further enhance the versatility of MAB algorithms by incorporating external information or 'context' into the decision-making process, enabling more nuanced strategies that adapt to dynamic environments and personalized requirements. This is particularly evident in applications such as personalized content recommendation and adaptive learning systems, where context plays a pivotal role in determining the optimal course of action. Similarly, Structured Bandits reveal underlying patterns in reward distributions, enabling a deeper understanding of the decision space and supporting more informed decision-making processes.

The practical implications of MAB algorithms are vast and varied, encompassing sectors as diverse as healthcare, where they optimize clinical trials and treatment strategies, to digital platforms, where

they refine content recommendation algorithms to enhance user engagement. In the educational domain, MAB algorithms personalize learning experiences, catering to individual student needs and preferences, thereby revolutionizing traditional pedagogical approaches. This paper endeavors to provide a comprehensive exploration of the intricate world of Multi-Armed Bandit algorithms, tracing their theoretical underpinnings, evolution, and the breadth of their applicability. Through this lens, we aim to underscore the transformative potential of MAB algorithms in navigating the complexities of decision-making across different realms, highlighting their role in driving innovation and efficiency in the face of uncertainty. As we delve into the nuances of various MAB strategies and their real-life applications, we illuminate the algorithms' capacity to not only solve complex problems but also to open new avenues for research and application in an ever-changing world.

2. Multi-armed Bandit Algorithms

MAB Algorithm is a set of methods aimed at maximizing rewards when faced with numerous choices. Depending on different scenarios and goals, various types of bandit algorithms have been developed.

2.1. Stochastic stationary Bandit

In the realm of Stochastic Stationary Bandits, the reward distribution for each action (or "arm") remains constant over time. This assumption simplifies the exploration-exploitation dilemma but still requires strategic action selection to maximize rewards. Early strategies like the ε-greedy algorithm balance this dilemma by primarily exploiting the best-known option while occasionally exploring at random. A more sophisticated approach is the Upper Confidence Bound (UCB) algorithm. It selects actions considering both the average rewards and the uncertainty or variance in those rewards, hence increasing the chance of discovering underestimated actions. The Thompson Sampling (TS) method takes a probabilistic approach, choosing actions based on the probability that they are optimal, given observed rewards. Each of these strategies aims to identify the most rewarding actions with minimal loss, striving for asymptotic optimality.

2.2. Restless Bandit

Restless Bandits operate under the premise that the reward distribution for each action changes over time. However, the change needs to be independent of the learner's actions. This dynamism introduces additional complexity, as it's no longer sufficient to identify the best action early on, which means continuous adaptation is essential. Algorithms in this category often incorporate mechanisms to detect and respond to changes in reward distributions, necessitating a balance between short-term gains and long-term adaptability.

2.3. Structured Bandit

In the problem of structured bandits, there exists a clear or implicit structure governing how rewards are allocated across different actions. Unlike traditional Multi-Armed Bandit (MAB) problems that focus on identifying the best arm to maximize rewards, structured bandit problems delve deeper into understanding the underlying reward structure. This understanding facilitates more informed decision-making, not just to maximize immediate rewards but also to unravel the reward mechanism itself, thereby optimizing long-term performance.

Linear Structured Bandits assume that the reward for each action is a linear function of hidden parameters. The core challenge in this process is to in accurately estimate these parameters to predict the rewards for different actions. Early decision-making may involve selecting actions that are highly informative about the parameters, even if they don't offer the highest immediate reward. In Linear Structured Bandits problems, even if it initially selects sub-optimal arms, better long-term optimization is still possible as the algorithm becomes confident about the parameter.

Global Structured Bandits deal with a more complex scenario where the reward structure can be non-linear and intricate and task is still to understand the structure to make informed decisions. Success

in global structured bandits depends on deciphering this structure to enhance decision-making and optimize rewards.

2.4. Contextual Bandi

Contextual Bandits extend the classic MAB framework by incorporating additional information or context that influences the decision-making process. In this enriched setting, the algorithm receives a set of contextual information before each action selection, which significantly impacts the reward outcomes. In contrast to the aforementioned algorithm, there no longer exists a universally optimal action. The task transforms into identifying the most suitable action within a specific context for each given contextual scenario. This framework is particularly suited to personalized recommendation systems, where the goal is to optimize actions not in isolation but as responses to specific situational contexts.

3. Practical Application

3.1. Medical & Healthcare

One of the earliest application for MAB is clinical trials. Clinical trials usually utilize MAB to recognize effective treatment and optimal dose.

In, the author proved the effectiveness of Multi-arm Bandit Problem in clinical trials though experiments and comparison with other classic allocation rules [1]. The research also summarized the design of clinical trials as an finite horizon case, and thus a restless bandit problem.

In the following years, a longitudinal series of researches about the modification of bandit algorithms on clinical trials were conducted. In the researcher use Asymptotically Optimal TS algorithm with an upper bound for the times selecting sub-optimal choice to deal with the finite time problem in clinic trials [2].

Instead of treating all the participants uniformly, contextual-bandit-based approaches are used to investigate precision oncology and Ischemic strokes treatment respectively, aiming at reducing participants'suffer by adapting the treatment based on their features [3].

Apart from clinical trials, the bandit algorithms are also used in other healthcare aspects like personal health-care assistant, public heath interventions and etc.

3.2. Content Recommendation

Content recommendation spans multiple fields such as advertisement recommendation, social media stream curation, news dissemination, and local service suggestions. Platforms in this space strive to personalize content for each user, aiming to maximize engagement and satisfaction. Crucially, the dynamic learning process inherent in content recommendation systems adapts to changing user preferences and continuously introduced new content.

The multi-armed bandit approach facilitates ongoing learning from user interactions, enabling platforms to adjust recommendations based on immediate feedback. Furthermore, users of content recommendation platforms display highly personalized characteristics; thus, leveraging user features as context can significantly enhance the overall reward of recommendations. For instance, the effectiveness of a contextual bandit method applied to social media streams recommendation was confirmed through experiments [4]. Particularly in advertisement recommendation, content recommendation intersects with communication studies. Introducing the concept of diffusion networks, this article employs the LinUCB algorithm, where spread serves as a reward, strategically targeting influential users to optimize overall dissemination [5]. Additionally, traditional MAB problems focus on selecting optimal arms, but practical communication challenges also require consideration of user attrition. To maximize total engagement, it is crucial to manage user retention and identify optimal arms. In response to the new challenge of MAB problems involving

abandonment, researchers proposed a method based on linear UCB theory that dynamically adjusts exploration intentions based on abandonment probabilities [6].

Another characteristic challenge in content recommendation is data sparsity. With content volumes nearing infinity, user data becomes sparse. One study addressed this by distinguishing nonresponses from negative responses before initiating the bandit process to better handle sparse data [7]. Employing knowledge graphs, another study used contextual information from user and item embeddings to design a knowledge-enhanced contextual bandit method that addresses both sparsity and dynamic constraints [8]. Further, several studies have explored the use of additional information as contextual data in recommendations. For example, one study extracted visual features from online advertisements to generate a quality ranking, which was then used in a hybrid bandit approach as supplementary context [9]. Another research acknowledged the value of follow-up information as a supplementary reward, modifying the contextual bandit approach in content recommendation platforms [10]. Unlike more specialized fields, content recommendation involves a comprehensive process that requires the integration of various elements. Considering the diverse components utilized in local service recommendation platforms like Yelp, a multifaceted contextual bandit algorithm was introduced and its effectiveness confirmed through experimentation [11].

3.3. Education

Course recommendation systems, often referred to as academic advising systems, are designed to personalize the learning experience by recommending courses based on a student's preferences, academic background, and career goals, which is the main practical application of bandit algorithm in education field.

Traditional course recommendation methods have primarily relied on filtering algorithms. However, bandit algorithms, although not mainstream, have also contributed significantly to the field.

The bandit algorithm is first used in off-line college course recommendation, serving as an academic assistance to promote students'school performance aims to minimize the time a student needs to graduate, taking into consideration factors like prerequisite requirements and course availability [12].

Online course recommendation systems is a broader field foe bandit algorithms. It has fewer constraints, such as major and graduation requirements, and offer a more personalized experience. The similarity between online course recommendations and content recommendations is notable; however, the unique characteristics of Massive Open Online Courses (MOOCs) have led to the development of many education-related online recommendation algorithms.

A significant challenge for MOOCs, which usually served as series, is the need for substantial storage resources addresses the issue of the immense storage cost. It proposes using distributed storage to manage courses and implements contextual bandits on an ever-expanding dataset to achieve dynamic and personalized MOOC recommendations [13].

Moreover, an essential aspect of education that must be considered in recommendations is the sequential nature of learning. Traditional contextual bandit models often overlook the importance of learning order, focusing on finding the most suitable courses without considering their sequence. This oversight can lead to inefficient learning pathways and wasted time. incorporates past student behaviors and current student states as context. This method gives full consideration to students' current knowledge bases, and also sequences the top-recommended courses in a logical order [14].

Additionally, integrates knowledge graph techniques to obtain feature vectors for MOOC courses, serving as the context for contextual bandits to enhance the effectiveness of recommendation [15].

3.4. Algorithm optimization

Beyond its practical applications in daily life, MAB have also made contributions to the research domain of computer science, notably in optimizing algorithms.

In the realm of machine learning, hyperparameter optimization serves as a fundamental tool for enhancing algorithmic performance. However the process demands intensive computational

resources. In, the authors employ the bandit algorithm to optimize the allocation of computational resources. By adjusting resource levels for evaluating hyperparameter settings, it reduces the computational burden, underlining bandit algorithms' potential to enhance model selection efficiency [16].

Investigates the Combined Algorithm Selection and Hyperparameter (CASH) optimization, aiming to streamline the selection of machine learning algorithms and their settings [17]. The proposed rising bandit solution effectively navigates the vast search space, illustrating the efficiency of bandit algorithms in automating and optimizing machine learning processes.

Besides, bandit algorithm is also used in network optimization explores optimizing data caching locations within networks to enhance performance [18]. By employing multi-armed bandit algorithms, the research dynamically optimizes caching strategies, showcasing bandit algorithms' adaptability in managing network resources under varying conditions.

4. Conclusion

This paper's exploration of Multi-Armed Bandit algorithms highlights their profound impact and versatility in tackling complex decision-making problems across various domains. From foundational strategies addressing Stochastic and Restless Bandits to the nuanced approaches of Contextual and Structured Bandits, we have observed the evolution of these algorithms to meet the needs of diverse, dynamic environments. The practical applications in sectors such as healthcare, content recommendation, and education underscore the significance and adaptability of these algorithms, demonstrating their potential to optimize outcomes in a range of contexts. The future trajectory of MAB algorithms points towards further refinement and the integration of advanced computational techniques, enhancing their capacity to comprehend and adapt to increasingly complex environments. This ongoing evolution not only promises to advance the theoretical foundations of MAB algorithms but also to broaden their applicability, offering more sophisticated, personalized solutions to the ever-changing challenges of decision-making in uncertain environments.

References

[1] Villar S S, Bowden J and Wason J. (2015). Multi-armed Bandit Models for the Optimal Design of Clinical Trials: Benefits and Challenges Stat Sci 30 199–215.

[2] Aziz M, Kaufmann E and Riviere M-K. (2021). On Multi-Armed Bandit Designs for Dose-Finding Trials Journal of Machine Learning Research 22 1–38.

[3] Varatharajah Y and Berry B. (2022). A Contextual-Bandit-Based Approach for Informed Decision-Making in Clinical Trials Life 12 1277.

[4] Gisselbrecht T, Lamprier S and Gallinari P. (2016). Dynamic Data Capture from Social Media Streams: A Contextual Bandit Approach Proceedings of the International AAAI Conference on Web and Social Media: 10, 131–40.

[5] Iacob A, Cautis B and Maniu S. (2022). Contextual Bandits for Advertising Campaigns: A Diffusion-Model Independent Approach Proceedings of the 2022 SIAM International Conference on Data Mining (SDM) Proceedings (Society for Industrial and Applied Mathematics). 513–21.

[6] Yang Z, Liu X and Ying L. (2024). Exploration, Exploitation, and Engagement in Multi-Armed Bandits with Abandonment Journal of Machine Learning Research, 25: 1–55.

[7] Yang S, Wang H, Zhang C and Gao Y. (2020). Contextual Bandits. With Hidden Features to Online Recommendation via Sparse Interactions IEEE Intell. Syst. 35: 62–72.

[8] Gan M and Kwon O-C 2022 A knowledge-enhanced contextual bandit approach for personalized recommendation in dynamic domains Knowledge-Based Systems 251 109158.

[9] Wang S, Liu Q, Ge T, Lian D and Zhang Z. (2021). A Hybrid Bandit Model with Visual Priors for Creative Ranking in Display Advertising Proceedings of the Web Conference 2021 WWW ’21, (New York, NY, USA: Association for Computing Machinery). 2324–34.

[10] Wang C, Ye Z, Feng Z, Badanidiyuru Varadaraja A and Xu H. (2023). Follow-ups Also Matter: Improving Contextual Bandits via Post-serving Contexts Advances in Neural Information Processing Systems, 36: 12774–96.

[11] Aziz M, Kaufmann E and Riviere M-K, (2021). On Multi-Armed Bandit Designs for Dose-Finding Trials Journal of Machine Learning Research. 22: 1–38.

[12] Xu J, Xing T and van der Schaar M, (2016). Personalized Course Sequence Recommendations IEEE Transactions on Signal Processing, 64: 5340–52.

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

[14] Intayoad W, Kamyod C and Temdee P. (2020). Reinforcement Learning Based on Contextual Bandits for Personalized Online Learning Recommendation Systems Wireless Pers Commun. 115, 2917–32.

[15] Ma D, Wang Y, Chen M and Shen J. (2021). SRACR: Semantic and Relationship-aware Online Course Recommendation 2021 IEEE International Conference on Engineering, Technology & Education (TALE) 2021 IEEE International Conference on Engineering, Technology & Education (TALE), 367–74.

[16] Sui G and Yu Y. (2020). Bayesian Contextual Bandits for Hyper Parameter Optimization IEEE Access, 8: 42971–9.

[17] Li Y, Jiang J, Gao J, Shao Y, Zhang C and Cui B. (2020). Efficient Automatic CASH via Rising Bandits, AAAI, 34: 4763–71.

[18] Tabei G, Ito Y, Kimura T and Hirata K. (2023). Design of Multi-Armed Bandit-Based Routing for in-Network Caching IEEE Access, 11, 82584–600.