Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/37/20230496

Investigation of progress and application related to Multi-Armed Bandit algorithms

Zizhuo Liu

Khoury College of Computer Science, Northeastern University, Boston, 02115, United States

liu.zizh@northeastern.edu

Abstract. This paper discusses four Multi-armed Bandit algorithms: Explore-then-Commit (ETC), Epsilon-Greedy, Upper Confidence Bound (UCB), and Thompson Sampling algorithm. ETC algorithm aims to spend the majority of rounds on the best arm, but it can lead to a suboptimal outcome if the environment changes rapidly. The Epsilon-Greedy algorithm is designed to explore and exploit simultaneously, while it often tries sub-optimal arm even after the algorithm finds the best arm. Thus, the Epsilon-Greedy algorithm performs well when the environment continuously changes. UCB algorithm is one of the most used Multi-armed Bandit algorithms because it can rapidly narrow the potential optimal decisions in a wide range of scenarios; however, the algorithm can be influenced by some specific pattern of reward distribution or noise presenting in the environment. Thompson Sampling algorithm is also one of the most common algorithms in the Multi-armed Bandit algorithm due to its simplicity, effectiveness, and adaptability to various reward distributions. The Thompson Sampling algorithm performs well in multiple scenarios because it explores and exploits simultaneously, but its variance is greater than the three algorithms mentioned above. Today, Multi-armed bandit algorithms are widely used in advertisement, health care, and website and app optimization. Finally, the Multi-armed Bandit algorithms are rapidly replacing the traditional algorithms; in the future, the advanced Multi-armed Bandit algorithm, contextual Multi-armed Bandit algorithm, will gradually replace the old one.

Keywords: Multi-Armed Bandit, ETC, UCB, Thompson Sampling, Epsilon-Greedy.

1. Introduction

The Multi-armed Bandit algorithm is a powerful and versatile approach used in decision-making processes that involve trade-offs between exploration and exploitation [1-3]. It is a classic problem inspired by the scenario of a gambler standing in front of multiple slot machines (arms) and trying to maximize their total reward over time. In the Multi-armed Bandit problem, each arm represents a choice or an action with an unknown reward distribution. The goal is to find the most rewarding arm while attempting to minimize the cumulative regret, which is the difference between the rewards obtained from the best arm and rewards obtained from selected arms over time.

Minimizing the cumulative regret requires a balance between exploration and exploitation. In the exploration phase, the algorithm seeks to find out the reward distribution of each arm by trying different actions and observing their outcomes. After the algorithm has gained sufficient information about the

DOI: 10.54254/2755-2721/37/20230496

arms, it starts exploiting the arm that found in exploration phase with the highest expected reward. This aims to maximize cumulative reward and reduces the regret over time. Since the algorithm will keep exploring the arm find in exploration phase with highest expected mean reward, it is essential to make sure that algorithm find the optimal arm in the exploration phase. However, spending more time on exploration can raise the cumulative regret, which is what algorithm trying to minimize. Therefore, balancing between exploration and exploitation is one of the main issue people considered.

The several advantages of Multi-arm Bandit algorithms are efficient resource allocation, continuous adaptation, and don't rely on history data. Multi-armed Bandit algorithms can allocate resources to different arms based on their performance dynamically, which can reduce resources wasted on suboptimal arm. Besides, the Multi-armed algorithm adapts to changing conditions and new data, allowing it to suit in a dynamic environment where preferences (rewards) might shift over time. Moreover, the algorithm can be used in wide areas because it does not require much historical data.

Furthermore, over time, several approaches have been developed to tackle the Multi-Arm Bandit problem, each with its unique exploration and exploitation strategies. Here will briefly discuss Explore Then Commit Algorithm (ETC) [4], Epsilon-Greedy Algorithm [5], Upper Confidence Bound Algorithm (UCB) [6], and Thompson Sampling Algorithm [7]. The ETC and Epsilon-Greedy Algorithms are the most basic, most straightforward, and earliest algorithms for solving multi-armed bandit problems, while the origin of the ETC and Epsilon-Greedy algorithms is not well documented. The UCB algorithm was first proposed in the early 2000s. The UCB1 algorithm, a specific instance of the UCB algorithm, was introduced by Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer in their paper, "Finite-time Analysis of the Multi-armed Bandit Problem" [8]. After the introduction of the UCB algorithm, it has become one of the foundational techniques in the field of reinforcement learning and has been extended and refined in various ways to address different variations of the Multi-armed Bandit problem and other exploration-exploitation challenges in machine learning. Finally, the Thompson Sampling algorithm, also known as the Bayesian Bandit algorithm, was first introduced by William R. Thompson in his paper, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples" (1933) [9].

Today, Multi-armed Bandit algorithms are widely used in machine learning, advertising, health care, and web page optimization field. According to Aman, Thompson Sampling algorithms are widely used in advertisements on multiple social media websites [10]. Moreover, UCB, Epsilon-greedy, and ETC algorithms are also widely used in advertising. For example, Li et al. created an algorithm by combining the Epsilon-greedy algorithm and the UCB algorithm in advertising, which perform much better than the traditional algorithm [11]. This paper will review different methods of Multi-armed Bandit algorithm in section 2. After reviewing each method, the application of Multi-armed bandit algorithms and their future will be further analyzed.

2. Methodology

2.1. ETC algorithm

The ETC algorithm employs a two-phase strategy, which are the Exploration phase and Commitment phase. During the Exploration phase, the algorithm uniformly explores each arm for a predetermined number of rounds denoted as 'k.' In this phase, the algorithm aims to gather information about the unknown reward distributions of all arms. After the algorithm finishing the exploration phase, it starts to apply the result obtained from the first phase. During the Commitment phase, the algorithm first identifies the arm that achieves the highest expected reward during exploration, then chooses that arm for the remaining rounds. The expected reward is the mean reward of each arm. Therefore, the ETC algorithm ensures that the majority of rounds are spent exploiting the best arm, maximizing cumulative rewards.

The ETC algorithm demonstrates effectiveness in scenarios characterized by elevated exploration costs or constrained opportunities for exploration. It proves advantageous when dealing with resource-intensive exploration or situations where extensive exploration is impractical due to constraints.

DOI: 10.54254/2755-2721/37/20230496

However, its performance might be suboptimal in cases where the reward distributions exhibit rapid changes. This limitation arises from the algorithm's commitment to a single arm after completing the exploration phase, without engaging in further investigation, which could potentially lead to suboptimal outcomes in rapidly evolving environments.

2.2. Epsilon-greedy algorithm

The Epsilon-Greedy Algorithm balances exploration and exploitation by employing a fixed exploration probability (epsilon, ε), where ε is an index that people set manually, and exploiting the current best arm with a probability of (1 - ε). The algorithm will first select each arm once to obtain their reward distribution information. After selecting each arm once, in the remaining round, the algorithm explores a randomly selected arm with a chance ε, and exploits the current best arm, highest mean reward, with a probability of (1 - ε). The algorithm ensures that all arms are given a chance to be pulled and exploit the best arm simultaneously.

The Epsilon-Greedy algorithm is characterized by its simplicity of implementation and has demonstrated effectiveness in various uncomplicated environments. Nonetheless, in situations where suboptimal choices persist over the long term, the algorithm may encounter challenges, primarily due to its tendency to continue exploration even after identifying the best arm. This prolonged exploration phase can hinder the algorithm's ability to focus on exploiting the arm with the highest expected reward, potentially leading to suboptimal performance in scenarios with sustained suboptimal choices.

2.3. UCB algorithm

The UCB algorithm applies the principle of optimism in the face of uncertainty to make decisions. The algorithm first runs each arm once to obtain their reward information. Then, it calculates the upper confidence bound for each arm. The UCB is derived from the mean reward of the arm and a confidence interval term that decreases with the number of times the arm has been pulled. At each round, the algorithm selects the arm with the highest UCB value, then updates that arm's UCB values. By following this, the UCB algorithm balances exploration by favoring arms with high uncertainty and exploitation by favoring arms with high expected rewards.

The UCB algorithm is renowned for its rigorous theoretical guarantees and its ability to accommodate diverse reward distributions. Its performance is commendable across a wide range of scenarios, and it exhibits rapid convergence towards the optimal arm. Nevertheless, the algorithm's efficacy can be influenced by specific types of noise present in the reward distributions, rendering it sensitive to such variations.

2.4. Thompson sampling algorithm

The Thompson Sampling algorithm employs a Bayesian approach, representing each arm's reward distribution with a probability distribution. Firstly, the algorithm selects each arm once to obtain their probability distribution. In each round, the algorithm first samples from each arm's posterior distributions and then selects the arm with the highest sampled value. After selecting the arm with the highest sampled value, the algorithm will update the probability distribution of the selected arm, then go to the next round. Therefore, Thompson Sampling effectively explores the arm by exploiting arms with higher probabilities of being optimal.

The Thompson Sampling algorithm has gained popularity due to its simplicity, effectiveness, and adaptability to various reward distributions. Moreover, Thompson Sampling performs well in both stochastic and adversarial environments because it explores and exploits simultaneously.

3. Applications and discussion

3.1. Advertisement

The Multi-armed Bandit algorithms are heavily used in choosing the best advertising strategies. Companies often need to decide how to allocate their budget among different advertisements or

DOI: 10.54254/2755-2721/37/20230496

marketing methods to increase user engagement, clicks, or conversions. Multi-armed Bandit algorithms help advertisers dynamically allocate resources to various alternatives (ads) depending on observed performance. More specifically, in advertising and marketing, the different advertisement options represent each arm, and the number of clicks or user engagement stands for the reward of each arm. Using Multi-armed Bandit algorithms, such as UCB and Thompson sampling algorithms, the cumulative regret of choosing sub-optimal arms has decreased significantly compared to traditional methods such as A/B testing.

Attaining optimal rewards demands a deliberate selection of algorithms by researchers, given the varied performance of these algorithms based on distinct situations. Additionally, certain algorithms can yield less-than-optimal choices, particularly in dynamic or highly turbulent environments. In the future, an increasing number of companies and organizations are poised to integrate Multi-armed Bandit algorithms into their advertising strategies, driven by their capability to curtail the expense associated with suboptimal decision-making.

3.2. Healthcare

Besides advertisements, Multi-armed Bandit algorithms are also prevalent in healthcare, especially in clinical trials. Multi-armed Bandit algorithms can assist doctors in determining the best treatment options for patients by analyzing patients' historical data. Applying Multi-armed Bandit algorithms in clinical trials, each arm represents different treatments, and the feedback from patients is the reward of each arm. Compared to traditional methods, the Multi-armed Bandit algorithms aim to maximize the chance of identifying the best treatment with minimal resources.

However, it's important to note that Multi-armed Bandit algorithms should be considered merely as guidelines due to the unique nature of patients' conditions. Furthermore, relying solely on the probabilistic perspective of the Multi-armed Bandit approach in medical treatments can be inherently unreliable. As a result, substantial manual intervention remains essential when applying Multi-armed Bandit algorithms in healthcare. Looking ahead, there is a growing inclination towards adopting contextual Multi-armed Bandit algorithms, which necessitate more personalized user information. This shift is driven by the highly individualized nature of healthcare, requiring insights into patients' medical histories, background conditions, and personal particulars. Thus, the contextual Multi-armed Bandit algorithm, especially combined with neural networks [12, 13], to be the evolving trend within the healthcare domain.

3.3. Website and app optimization

Multi-armed Bandit algorithms are critical in enhancing user experience and engagement on websites and mobile applications. Website and app developers often need to decide the web page's design, content, and placement. In the practice of Multi-armed Bandit algorithms in website and app optimization, different webpage designs can be represented as each arm, and the feedback, such as click rate and user rating, can be treated as the reward of the arms. Compared to traditional algorithms, such as A/B testing, Multi-armed Bandit algorithms can effectively allocate the resources to each design to pick the optimal configuration for the website and app.

Nonetheless, Multi-armed Bandit algorithms carry the inherent risk of occasionally leading to suboptimal webpage strategies due to the fluidity of people's preferences over time. Consequently, developers must frequently recalibrate the weighting of each factor to ensure these algorithms converge on the optimal choice. Looking ahead to the next two decades, the prevalence of Multi-armed Bandit algorithms is projected to surge across various optimization domains. As the world's trajectory propels further into the digital era, the efficiency of discerning optimal decisions gains paramount importance. In contrast to conventional approaches, Multi-armed Bandit algorithms offer a resource-efficient avenue for exploring suboptimal options, thus solidifying their position as the forthcoming trend in website optimization.

DOI: 10.54254/2755-2721/37/20230496

4. Conclusion

This paper briefly introduces the history, principle, and function of Multi-armed Bandit algorithms and their application. The four Multi-armed Bandit algorithms discussed in the report are ETC, UCB, Epsilon-Greedy, and Thompson Sampling algorithm. Today, Multi-armed Bandit algorithms are widely used in advertisements, healthcare, and website and app optimization because they can significantly improve efficiency and reduce the cost of choosing sub-optimal decisions. The review's drawback lies in the insufficiency of direct evidence and data to substantiate the projections regarding the future utilization of Multi-armed Bandit algorithms. Subsequently, forthcoming iterations of the paper are anticipated to delve into the realm of contextual Multi-armed Bandit algorithms, an advanced iteration of the algorithm that leverages richer personal and environmental data to pinpoint optimal solutions. This trajectory is driven by the algorithm's capacity to furnish decisions that are both more personalized and precise, thereby enhancing its efficacy.

References

[1] Vermorel J Mohri M 2005 Multi-armed bandit algorithms and empirical evaluation European conference on machine learning. Berlin, Heidelberg: Springer Berlin Heidelberg 437-448

[2] Kuleshov V Precup D 2014 Algorithms for multi-armed bandit problems arXiv preprint arXiv:1402.6028

[3] Slivkins A 2019 Introduction to multi-armed bandits Foundations and Trends® in Machine Learning 12(1-2): 1-286.

[4] Nie G Agarwal M Umrawal A K et al 2022 An explore-then-commit algorithm for submodular maximization under full-bandit feedback Uncertainty in Artificial Intelligence. PMLR, 1541-1551

[5] 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) IEEE 929-936

[6] Garivier A Moulines E 2011 On upper-confidence bound policies for switching bandit problems International Conference on Algorithmic Learning Theory. Berlin, Heidelberg: Springer Berlin Heidelberg 174-188

[7] Russo D J Van Roy B Kazerouni A et al 2018 A tutorial on thompson sampling Foundations and Trends® in Machine Learning 11(1): 1-96

[8] Auer P Cesa-Bianchi N & Fischer P 2002 Finite-time Analysis of the Multiarmed Bandit Problem Machine Learning 47, 235–256 https://doi.org/10.1023/A:1013689704352

[9] 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

[10] Aman A 2021 Thompson sampling in social media marketing Towards Data Science https://towardsdatascience.com/thompson-sampling-in-social-media-marketing-97d1892b125f

[11] Li W et al 2010 Exploitation and exploration in a performance based contextual advertising system Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining https://doi.org/10.1145/1835804.1835811

[12] Qiu Y Wang J Jin Z et al 2022 Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training Biomedical Signal Processing and Control 72: 103323

[13] Al-Shayea Q K 2011 Artificial neural networks in medical diagnosis International Journal of Computer Science Issues 8(2): 150-154