Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/34/20230346

The investigation and application of muti-armed algorithms in recommendation systems

Xiaoning Zhang

The Department of Maths and Statistics, the University of Connecticut, Mansfield, 06250, the United States

Xiaoning.zhang@uconn.edu

Abstract. In the contemporary Internet recommendation systems in various fields, muti-armed algorithms will show high accuracy and practicability. Internet users can always get abundant positive feedback through the recommendation system via utilizing these algorithms. In this paper, there are three most typical muti-armed algorithms as examples are provided. Explore-Then-Commit (ETC) algorithm is the first to mentioned. The physical meaning of this algorithm is that in the exploration stage, the action will be selected in a certain order, and after a certain number of rounds, the action with the largest average reward will be directly selected. Moreover, Upper Confidence Bond (UCB) algorithm is also a type of pivotal tool. The main function of UCB algorithm is to make the selection by calculating the upper bound of each arm confidence interval instead of the expected reward of the slot machine. Thus, it is an optimistic algorithm. First, select each arm in random order, then calculate the value of the upper bound of each arm confidence interval, and finally, select the arm with the largest value. The last mentioned Thompson Sampling (TS) Algorithm take Bayesian optimization as the theoretical basis. This algorithm takes out the candidate parameters, generates a random number, and selects the maximum value for input. This paper also introduced two fields related to the topic of the application of muti-armed algorithms in the recommendation systems in modern fields. News recommendation algorithms and personalized recommendation systems can be more comprehensive and representative to illustrate the practicality of these algorithms. Therefore, muti-armed algorithms reflect the importance in the recommendation fields via the balance of exploration and exploitation.

Keywords: muti-armed algorithms, recommendation fields, exploration.

1. Introduction

The Multi-Armed bandit problem, rooted in the gambling domain involves a scenario where multiple slot machines, outwardly indistinguishable but with varying winning probabilities, are available to gamblers. Given limited information about the winning probabilities and a constrained number of attempts, the challenge lies in determining how to select the slot machines in a manner that maximizes the overall return. Therefore, this problem centers on making a decision in order to obtain the "maximum reward" in the case of a limited time of attempts and uncertainty situation. There is no doubt that the problem of the algorithms of Multi-Armed bandit problem has significant importance due to their wide-ranging applicability in various domain such as the short video recommendation.

DOI: 10.54254/2755-2721/34/20230346

In 1933, bandit problems were initially introduced in the study. After that, in the 1950s, Mostler et al. conducted a series of experiments on mice [1]. The experimental paradigm involved placing the mouse at the bottom of a T-shaped maze and presenting it with a decision-making task to choose a direction in order to access food. The mouse was unaware of the location of the food endpoint in advance, thus confronting the dilemma of selecting either the left or right path to obtain the food reward. Apply this to humans, it is similar with a 'two-armed bandit' machine. Human can choose to pull either the left or the right arm of the machine, in order to get a random benefit for each arm of this machine. This is also the origin pf the name bandit problems [1]. Because the algorithms of this kind of problem is very practical, it is widely used in various fields. For example, these algorithms are often used by various companies to analyze data and then optimize product prices according to users' data. Also, the most classic use is to analyze abundant kinds of questions about gambling.

Currently, the Muti-armed bandit algorithms are applied on more new technology fields. Compared with the applications in the traditional fields mentioned above, the muti-armed algorithm is more widely used in the contemporary new fields. Due to the abundant increasing number of digital information and Internet users, accessing items of interest may be a potential problem. Recommendation system can filter out qualified information from massive digital information according to users' preferences and demand, so as to effectively solve the problem of information overload [2]. This paper will provide an overview of this system of field. The research topic introduced on this paper will focus on a large number of media websites or apps push videos, advertisement or news to specific user according on some database and the use of muti-armed algorithms. Muti-armed algorithms play a pivotal role in the operation mechanism on how to attract users with content of these online media websites, which allow Internet users to publish, share and watch audio, video or article works smoothly. Thus, conduct the research on this can be very helpful on learning and researching muti-armed problems and related algorithms more efficiently.

The rest of this paper included several sections about the operation principle and methods of muti-armed bandit algorithms served in accurate push of results on video or short video media website or apps. This paper will research and explain them in some narrow respects and examples. The section 2 will introduce some classic muti-armed bandit algorithms in recommendation system field. The section 3 will apply some actual scene and discuss which tasks are these algorithms applied to. The section 4 will draw a conclusion and summarize the full text and look forward to the future.

2. Method

2.1. Explore-then-commit algorithm

One of the foundational algorithms introduced in this study in the realm of MAB is the Explore-Then-Commit Algorithm (ETC). A critical aspect of the ETC algorithm revolves around the determination of the number of experiments to be conducted on each arm, denoted by a natural number "m." If there are k actions, this algorithm will explore for mk rounds before selecting a specific action for the rest of the round. The policy of ETC algorithm is imputing m firstly, and then choosing an action below. If t less than or equal to mk, A(t) may equal to (tmod+1), and A(t) will equal to argmax$i\mu$i(mk) if t greater than mk [3]. ETC algorithm is contributed with exploration phase and exploitation phase. These two phases complement each other. The vanilla ETC strategy can bring some concrete steps to operate. In the first step, the main task is exploration. The mode of operation is pulling all arms for the same number of times, and this number may be a fixed value based on a stopping time according to the data. In the next step, called exploitation stage, based on the outcome on the last step, the arm that achieves the best average reward will naturally be pulled [4]. However, despite its pragmatic approach, the ETC algorithm suffers from a notable drawback. In terms of asymptotic optimality, it does not enjoy a first-priority status, as achieving asymptotic optimality proves relatively challenging. Consequently, in the recommendation domain, the ETC algorithm may not outperform the Upper Confidence Bound Algorithm.

DOI: 10.54254/2755-2721/34/20230346

2.2. Upper confidence bound algorithm

The Upper Confidence Bound (UCB) Algorithm introduced represents a prominent and practical alternative to the Explore-Then-Commit Algorithm. UCB algorithm is an algorithm that does not consider randomness completely. It not only considers the reward, but also considers the confidence value of the reward. There are n stochastic arms with unknown distributions. Based on this, the operation is pulling an arm to get a reward from the reward distribution in each time slot. Additionally, the goal of this algorithm is maximizing the cumulative reward and minimizing the regret [5]. Compared with ETC algorithm, UCB algorithm has several advantages. It can operate more effectively and accurately than the ETC algorithm when there are two or more arms. Moreover, it depends less on professional advance knowledge of the aspect about suboptimality gaps [6]. Therefore, in the recommendation system, UCB algorithm can get more benefits, avoid regret and integrate various factors, so as to screen out the decision-making algorithm suitable for the current environment. An illustrative example of this is the adversarial bandit problem, which involves a gambler striving to achieve the highest total reward by selecting which slot machines to play out of "K" machines. To optimize his reward, the gambler faces the exploration-exploitation trade-off. Continuously playing the same slot machine represents exploitation, while attempting different machines constitutes exploration. [7]. In other words, the central idea of the UCB algorithm is that since the current arm with high success rate has high utilization value, and the arm with high uncertainty has high exploration value, the algorithm wants to integrate these two values and give each arm an evaluation, so as to select the arm with the highest value. Adversarial bandit problem makes no statistical assumptions, so the reward may be difficult to generate by traditional algorithms. Therefore, this kind of new algorithm is developed, which uses upper confidence bounds to calculate the evaluated reward. UCB algorithm can effectively controls regret fluctuations using the upper bound of confidence [8].

2.3. Thompson sampling algorithm

The third algorithm is Thompson Sampling (TS) algorithm. The main idea of TS algorithm is to set the reward distribution parameters of each arm to have a simple prior distribution, and operate an arm based on calculating the probability of being the most desirable arm.after any time step [9]. The mathematical theory of the algorithm is based on Bayesian optimization. Bayesian optimization through beginning with the prior belief distribution of the function, and then merges the function evaluation into the updated belief in the form of a posteriori. In this process, numerous algorithms evaluate the value of this function by querying some specific rules. Then, randomly select a program that uses rear-maximized random samples to select a point and call it Thompson sampling [10]. Thus, the TS algorithm is a method to analyze the circumstance and reward distribution of each arm in terms of random probability and make a final decision. It randomly selects each arm according to the optimal probability of each arm and addressing the exploration/exploitation trade-off based on the observed facts. Additionally, the difference between TS algorithm and UCB algorithm is that UCB algorithm chooses the arm with the highest upper confidence, while TS algorithm chooses arm according to random probability.

3. Applications and discussion

3.1. News recommendation algorithm

Currently, a substantial number of news websites or applications embed some muti-armed bandit algorithms to assists users in filtering and presenting news that may be of interest from a giant number of news, which can undoubtedly effectively improve users' experience of reading news [11]. There into, the most direct to achieve the goal is the UCB algorithm. Since the items in a vast amount of news information generally vary greatly, UCB may explore many items that have not been explored. However, such random exploration can lead to time-consuming and recommendation performance problems, so the more advanced Dueling Bandit Gradient Descent algorithm is practiced exploring [12]. The primary function of this algorithm is to determine the minimum extreme value of a function.

DOI: 10.54254/2755-2721/34/20230346

The general process is the agent utilizes the current network N to generate the recommendation list R, and then employs another network N' to generate another list R'. The next step is for the agent to combine R and R 'and generate a new recommendation list L via adding interference formulas to the current network N and probability interleaving. If the agent selects R, put the items from R into L and show them priority. Subsequently, L will be recommended to users. If the agent obtains a satisfied feedback, keep the current network N unchanged. If the items from R' gets better feedback, the agent will update the current network to N'. It is more efficient and convenient to explore in this method [12].

3.2. Online personalized recommendation systems

Additionally, muti-armed algorithms are also particularly practical in a variety of online personalized recommendation systems in different online stores. These personalized recommendation systems that applying in online stores will be very efficient in recommending appropriate items to consumers, according to the specific information of consumers and products. Although these systems face some challenge, for example, the cold start problem. The recommendation system may not take effect on brand-new users or projects, or the popularity of some products may be timely [13]. This is also a classic exploration/exploitation dilemma. In this scenario, contextual multi-armed bandit is widely utilized. The online algorithm in focus is designed to generate queries continuously, with each query comprising a comprehensive history of past queries pertaining to the item, encompassing information such as the number of queries and the inquirer's details. It clusters the context of the query into similar regions and runs it separately for each context set. As a result, this algorithm naturally captures online scenes over time, and in this method, it is widely used in personalized recommendation services to solve the cold start problem mentioned above [14]. In addition to the commodity personalized recommendation system, advertising push system and web page optimization system are also fields in which this algorithm is often very practical.

4. Conclusion

This paper undertakes an investigation into multiple multi-armed bandit algorithms and their application within specific contexts of the recommendation domain. The utilization of multi-armed bandit algorithms in this context is characterized by a swift and comprehensive evolution. Analysis of the current advancement within this field reveals the widespread integration of multi-armed bandit algorithms into recommendation systems and related domains. Nonetheless, it is important to acknowledge that the extant exploration of the algorithms' application within recommendation systems lacks thoroughness, resulting in potentially inadequately detailed or accurate depictions of their practical implementation. Future progress hinges upon a deeper scrutiny of the recommendation system domain and a more comprehensive understanding of these multi-armed bandit algorithms, serving as the linchpin for advancing research endeavors.

References

[1] Bandit Algorithms Tor Lattimore and Csaba Szepesv ari chapter 1 pp 8
[2] Isinkaye F O Folajimi Y O & Ojokoh B A 2015 Recommendation systems: Principles, methods and evaluation Egyptian informatics journal 16(3) 261-273
[3] Bandit Algorithms Tor Lattimore and Csaba Szepesv ari chapter 6 pp 90–99
[4] Jin T Xu P Xiao X & Gu Q 2021 July Double explore-then-commit: Asymptotic optimality and beyond In Conference on Learning Theory (pp 2584-2633) PMLR
[5] Li J 2016 Muti-armed Bandits Online Learning and Sequential Prediction
[6] Bandit Algorithms Tor Lattimore and Csaba Szepesv ari chapter 7 pp 100–115
[7] Auer P Cesa-Bianchi N Freund Y & Schapire R E 1995 Gambling in a rigged casino: The adversarial multi-armed bandit problem In Proceedings of IEEE 36th annual foundations of computer science (pp 322-331) IEEE
[8] Auer P 2000 Using upper confidence bounds for online learning. In Proceedings 41st Annual Symposium on Foundations of Computer Science (pp 270-279) IEEE

DOI: 10.54254/2755-2721/34/20230346

[9] Agrawal S & Goyal N 2012 Analysis of thompson sampling for the multi-armed bandit problem In Conference on learning theory (pp 39-1) JMLR Workshop and Conference Proceedings
[10] Kandasamy K Krishnamurthy A Schneider J & Póczos B 2018 Parallelised Bayesian optimisation via Thompson sampling In International Conference on Artificial Intelligence and Statistics (pp 133-142) PMLR
[11] Tian X Ding Q Liao Z & Sun G 2021 A review of news recommendation algorithms based on deep learning Computer Science and Exploration 15 (6) 971
[12] Zheng G Zhang F Zheng Z Xiang Y Yuan N J Xie X & Li Z 2018 DRN: A deep reinforcement learning framework for news recommendation In Proceedings of the 2018 world wide web conference (pp 167-176)
[13] Zeng C Wang Q Mokhtari S & Li T 2016 Online context-aware recommendation with time varying multi-armed bandit In Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining (pp 2025-2034)
[14] Lu T Pál D Pál M 2010 Contextual multi-armed bandits Proceedings of the Thirteenth international conference on Artificial Intelligence and Statistics. JMLR Workshop and Conference Proceedings 485-492