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

Optimizing Reward and Minimizing Comprehensive Cost in Recommendation Systems

Shuohan Gu*

The economic and management college, Communication university of china, Beijing, China

* Corresponding Author Email: 2020216083002@cuc. edu.cn

Abstract. An in-depth discussion will explore the advanced methodologies designed to minimize regret in recommendation systems, shedding light on the principal strategies and assessing their effectiveness. This analysis will also consider additional factors that could profoundly impact the decision-making mechanisms within these systems. Special focus will be placed on user-specific variables such as contextual nuances, temporal dynamics, and individual preferences that, although frequently overlooked, have the potential to significantly improve the personalization and efficacy of recommendations. By incorporating these elements, the goal is to enhance the adaptability and predictive accuracy of recommendation systems, ultimately leading to a more engaging and satisfying user experience. This progressive approach is set to redefine the standards for user satisfaction and system efficiency in the dynamic realm of digital interactions and personalized technology, marking a significant evolution in how these systems cater to individual needs and preferences.

Keywords: Optimizing Reward; Minimizing; Comprehensive.

1. Introduction

In the realm of recommendation systems, the multi-armed bandit (MAB) framework has emerged as a pivotal model for optimizing user interactions and maximizing rewards. This probabilistic approach effectively addresses the inherent uncertainties in user preferences and item attributes, making it indispensable in applications like online recommendation systems where decision-making must be both swift and accurate. The fundamental challenge within the MAB framework lies in the dual necessities of exploration and exploitation. Exploration requires the algorithm to investigate various options to accrue valuable data about their potential, while exploitation focuses on leveraging the most promising options based on accumulated knowledge to maximize immediate rewards.

The significance of this balance cannot be overstated, as it directly influences the system's ability to recommend items that are both novel and aligned with user preferences, thereby enhancing user satisfaction and engagement. Furthermore, the MAB model is particularly adept at tackling various forms of the cold start problem—whether it pertains to new users, new items, or new systems—by enabling the algorithm to learn rapidly from interactions despite the initial lack of data. Current research in bandit algorithm strategies not only continues to refine the traditional balance between exploration and exploitation but also introduces sophisticated mechanisms to minimize regret, thus indirectly maximizing the cumulative reward over time. By incorporating advanced techniques such as Upper Confidence Bound (UCB) and Thompson Sampling, bandit algorithms improve their decision-making processes, making them more dynamic and responsive to changes in user behavior and item popularity. This introduction sets the stage for a deeper exploration of how these methodologies can be further enhanced by considering additional user-centric variables such as contextual influences, temporal dynamics, and personal preferences. By integrating these factors, the aim is to propel the adaptability and predictive accuracy of recommendation systems to new heights, thereby redefining the benchmarks for user satisfaction and system efficiency in the rapidly evolving landscape of digital interaction and personalized technology.

2. Mainstream Objectives in Bandit Algorithm Research

2.1. Exploration and Exploitation: Balancing Phases

During the bandit algorithm, the balancing between the exploration and exploitation phases in order to reach the relatively optimal result has been studied since the bandit algorithm had been raised as a practical method, which is also called exploration-exploitation trade-off [1]. During the exploration phase, the capacity to insight the valid information needs to be maximized;similarly, during the exploitation phase, the capacity of the usage of the valid information needs to be maximized [2].

The utilization of the bandit algorithm in the recommending system has exerted positive effects as below:Firstly, it solves the item cold start problem(including user cold start (no historical behavior data after new user registration), item cold start (no user interaction data for newly added items) and system cold start (it is recommended that all relevant historical data is lacking when the system is launched)) [3]. The exploration and exploitation algorithm can have a probabilistic tendency for the newly added items, make the items pass the cold start stage quickly through the exposure of the items and user feedback, quickly find the potential items, and enrich the collection of high-quality items [4]. Furthermore, it explores new points of interest for users. Since users' interests will shift over time, exploring new items with a certain probability can also test users' interest and preference for new items, timely capture the trend of users' interest and preference transfer, and enhance users' stickiness to the platform [5]. Additionally, it exposes the richness of the item. By adding a part of the explored new items to the items with strong personalized recall, the diversity of display results can be increased, and a large number of homogenized items can reduce the user's visual fatigue. Exposure of the richness of items, by adding a part of the exploration of new items in the personalized recall of strong items, can increase the diversity of display results, reduce a large number of homogenized items users produce visual fatigue [6].

2.2. Reducing Regret in Recommendation Systems

To maximize rewards in a multi-armed bandit setup, an optimal balance between exploration and exploitation must be achieved. Exploration involves sampling from all available arms to gain more information about their reward distributions [7]. The value of regret is a Conversely, exploitation involves selecting the arm that currently has the highest estimated value based on the information gathered. Starting from the most classical naive algorithm, we will discuss the operation mechanism of the algorithm itself and the regret value.

The UCB approach is optimistic in the face of uncertainty and tends to naturally decrease exploration over time as it becomes more confident in its estimates. The UCB based algorithms with complexly set indexes were raised and the log linear regret bounds were developed for the algorithm [8].

Another classical method is the Thompson Sampling (TS), which leverages a probability matching strategy. It samples from the posterior distribution of each arm's rewards and then chooses the arm with the highest sample [9]. This Bayesian approach dynamically adjusts the level of exploration based on the obtained rewards and is particularly effective when rewards are stochastic.

By minimizing regret, the algorithm indirectly maximizes the expected total reward over the course of the interactions. "

Both algorithms have the ability to reach the optimal result during the exploitation-exploration phase, however, as context-free Bandit algorithms, they just focus on how to make the choice instead of looking at what kind of arm it's dealing with [10]. Thus it is relatively hard to reach “appreciable predictive validity” in many occasions.

3. Active Reinforcement Learning: Minimizing Total Cost

The main goal of the reinforcement learning is to maximize the net rewards, that is not only related to the total reward, but also associated with the total cost.

3.1. Agent, State, and Action: Learning Based on Conditions

3.1.1. Agent:Different learning performance in experts and non-experts.

In reality, the situation of the recommending system we often met is :Someone goes online and sees different sorts of online advertisements and makes his own choice, so the result remains unknown before the clicking-action. We must realize that the recommending effects is, to some extent, related to the different learning abilities of the agents——expert and non-expert.

Jaron T. Colas, John P. O'Doherty, Scott T. Grafton have developed a model GRL, which, comparing with the common reinforcement learning, paying more attention to the different Whether the agent is embodied as robot or human, it learns from feedback to make decisions and select physical actions that maximize future reward while minimizing various costs. Proofed by the experiments above, to some extent, the forces of bias and hysteresis antagonize and compete with the different performance in learning , a negative correlation was discovered between learning performance and the weight of action bias and hysteresis.

3.1.2. State:related to the specific situation like contextual ones.

The state of the reinforcement learning is related to the specific place and moment, that is, a specific instant configuration. An algorithm named Contextual bandit is put forward and studied.

By applying the online linear classifier, we can reach the similar results by ultilizing the algorithm below:

Contextual-Epsilon-greedy strategy: They can be adjusted by determining which situation is most relevant to the exploration phase or the exploitation phase, resulting in high exploration behavior in non-critical situations and high exploitation behavior in critical situations.

3.1.3. Action:Vaious elements jointly make the decision.

An Action is a round of choice that an agent can take. An action is almost obvious, but it should be noted that the agent is choosing from a list of possible actions. Various factors jointly contribute to the selection in every round.

The most relevant factor in the Choice-making is related to the algorithm. The different algorithms make its unique decisions according to the operation mechanism. For instance, the attitude of UCB algorithm in decision-making is optimistic when estimating rewards: setting rewards at the upper end of the confidence interval;while the the attitude of Thompson Sampling algorithm is heuristic with the prior experiences, typically a Bayesian approach when the prior belief is updated referencing to the data in the exploration phase and form its own posterior distribution.

In addition, it is non-negligible that the effect of learning capability and mechanism have prominent influences on the decision-making. Due to the different learning effects of the individual agent, the decision-making of the action varies a lot.

3.2. The practical strategies with the potential capability of minimising the total cost in recommending system

Firstly, Bandit algorithms can update recommendations in real-time as a user interacts with the system, iteratively improving through a continuous process of exploration and exploitation, which indicates that the system can adjust its recommendation strategies quickly based on immediate user feedback, increasing metrics like click-through rates.

With Bandit algorithms, recommendation systems can converge to the optimal or suboptimal recommendations quickly after a small number of interaction trials, which presenting a vast array of options to gauge user preferences is not necessary, thus saving on recommendation costs and reducing user decision fatigue.

Contextual Bandits can leverage additional environmental or user information to provide more accurate recommendations, including factors like time, location, device, or user's historical behavior, making the recommendations more personalized.

4. Methodology

We raise an active inference framework to reduce inference and decision costs with a math model. According to K. Gokcesu and S. S. Kozat, the adversarial multi-armed bandit problem can be solved due to a math framework.

4.1. The description and expansion of the math framework

First, the framework would be established with the necessary variations.

The adversarial multi-armed bandit problem is set with the number of the bandit arms is set as M and one of the arms at each round is set as T in random.

A valid manner of quantifying the hardness of learning the optimum strategy is the number of switches Sn it has as the column vector, due to the fact that every time the selection switches, the action of learning is repeated the optimal arm from scratch. If the optimal strategy has S−1 switches in its arm selection, we would believe that the optimal strategy has S segments.

In addition, S can be plot as the total switches as below:

$$S = 1 + \sum_{t=2}^{T} 1_{s_n \neq s_{n-1}} \tag{1}$$

The selection of the user in calculations expressed as the column vector can be represented by ut as below:

$$u_t = u_t(l_{u_{t-1}}; u_{t-1}) \quad u_t \in \{1, \ldots, M\} \tag{2}$$

The accumulated loss L at time T of any strategy Sn can be described as below:

$$L_{S_T} = \sum_{t=1}^{T} l_{t,s_t} \tag{3}$$

The regret Rn is highly related to the hardness it has to reach the level of the optimum strategy. Rn is highly related to the s, as described as below:

$$R_T = \sum_{t=1}^{T} l_{t,u_t} - \sum_{t=1}^{T} l_{t,s_{t^*}} = L_{u_t} - L_{s_{t^*}} \tag{4}$$

Hence, we define the optimal strategy without assumptions of the loss, the performance is given as below:

$$S_T^* = arg_{S_T} min L_{S_T} \quad 1 \leq t \leq T \tag{5}$$

Hence, we design the weight w_{s_t} which shows our reliability on the weight-based strategy and combination weights, $\Gamma(\cdot)$, in a way that its update at each round t is independent from everything except whether a switch is made or not. Hence, the two variations are to be discussed and calculated to make efforts to approach the optimum reward.

4.2. Reach the optimal selection of "mini-max strategy"

Hence, our goal is to introduce an algorithm that reach the mini-max optimal expected regret. According to K. Gokcesu and S. S. Kozat, through a serious of the math theoretical inference, a theorem with the maximum bound of logarithmic terms is given, who successfully minimize the online regret:Deducing the upper bound of logarithmic terms with the expression of $E[Rn] \leq \check{O}(\sqrt{MST})$, whether or not the number of switches S and the round T are known. When algorithm weights are uniform, the combined outcome yields a regret of O(TlogN)^1/2 and operates with O(N) computational cost, given N as the count of unified algorithms. Therefore, merging an exponential quantity of these algorithms without strategy leads to persistent regret, marked by O(T), and needs time-intensified exponential computational effort. It is thus imperative to smartly fuse these tactics by selecting combination weights that are both strategically efficient and computationally feasible in real-time, aiming for diminishing regret within a polynomial time frame.

As is discussed above, the equality of the weight has a relatively higher standard regret. In order to achieve the mini-max optimal regret, hence, the weight should be assigned properly. In the second component, the previous performance of st is relied on entirely, with the figure of $\exp(-\eta \acute{L}_{S_t(1:t-1)})$. Thus, the combined weight can be given as:

$\omega_{s_t} = T(s_t)e^{-\eta \acute{L}_{S_t(1:t-1)}}$, Where η is the learning rate and $\acute{L}_{S_t(1:t-1)}$ is the unbiased estimator of $L_{S_t(1:t-1)}$.

The telescoping rule have been scheduled is T(st)=T(st|st(1:t−1))T(st(1:t−1)), hence $\sum_{m=1}^{M} T([s_t: m]|s_t) = 1$, ∀st, t∈{0, …, T−1}. Thus, an expression of the expected regret satisfying the formula above yields:

$$E[R_T] \leqq \min_{S_T}\left(\frac{\eta MT}{2} + \frac{1}{\eta}\ln W(st) + L_{s_t} - L_{s_{t^*}}\right) \quad (6)$$

The implication of the expected regret is that with a judicious configuration of $T(s_t)$ and η, it's possible to attain regret that is not only sub-linear but can also be minimized optimally. Nevertheless, the method for assigning weights to $T(s_t)$ must allow for sequential construction and must also adhere to certain criteria.

4.3. The valid translation of the optimal strategy in the recommending system

4.3.1. The application in reality--minimizing the regret and loss in the online recommending system.

For the algorithm mentioned in 4. 1&4. 2 is empirical in almost all the sort of big data applications, the most extensive use——the recommending system has the hugest range of applications that can be applied. As is mentioned above, when the weights of the exploration in the M round take the uniform value, the regret value may rise to the standard of O(TlogN)^1/2 , in addition, with extra computational effort. Hence, here comes the question that the setting value of the parameter ω, with the value of the figure S and T rarely exert prominent influences due to the online designing idea.

Hence, in practical recommending process, the decision of the weight contributes a lot to the effect of the minimizing the value of the regret. The target users of the cold start problem can be divided into three types: new visiting users or new registered users. Users whose behavior track is very sparse and whose personal information is basically absent; Users in industries related to low-frequency application scenarios. For these target users in these three scenarios, the main strategy of cold start mainly includes two aspects: fully mining the user's personal information and making full use of existing resource information, the former is about the exploration phase while the later is about the exploitation phase. Considering the former problem, when the exploration phase meets the cold start problem and during the early stage of the exploring, the decision of the weight ω must be carefully

determined:Initially, the value S between the st and ut is undoubtedly large, and the regret $R_T = \sum_{t=1}^{T} l_{t,u_t} - \sum_{t=1}^{T} l_{t,s_{t^*}} = L_{u_t} - L_{s_{t^*}}$ would be huge. Due to the formula $E[R_T] \leqq \min_{S_T}(\frac{\eta MT}{2} + \frac{1}{\eta}\ln W(st) + L_{s_t} - L_{s_{t^*}})$, when the value of t、$L_{s_t} - L_{s_{t^*}}$ and η are fixed, and the weight ω is mainly determined by a-priori weight T(st). To minimize the regret R_T, the T(st) should be low.

However, the aim of every recommending system is to insight as many preferences as possible of the users with the fewest t. We assume that the t is fixed, to balance the aim of minimizing R_T and the need of insight, the T(st) need to be set at a certain value, thus the value of $L_{s_t} - L_{s_{t^*}}$ should be minimized according to the user's basic label to recommend or simply directly recommend the popular or manual operation to recommend.

4.3.2. The actual condition--the limited budget in the online recommending system.

In the realistic application of the bandit algorithm in the recommending system, the most common situation we must face is that the company's resources are limited, that is, the budget of the exploration-exploitation phase must be minimized and maitain at a certain level. Hence, the budgeted MAB is proposed.

Due to the essence of the bandit algorithm——dealing with the balance between the exploration-exploitation trade-off, some measures have been taken into practice in order to approach this. Taking the UCB algorithm as an instance, the algorithm i-UCB is put forward, in which during the exploitation phase, the assessment criteria is adjusted from the reward to the reward-to-cost ratio.

Thus, in the budget-limiting situation, the recommending system may put more attention on the cost and transform ratio from cost to reward. In reality, the budget of the exploration is the key consideration of the company.

5. Results and Discussion

5.1. Results:the optimum application in recommending system

Based on the algorithm and the analysis above, the optimum application in the recommending system should be like this:

When the recommending system met the cold start problem, according to the mini-max optimal algorithm, the optimal choice is to minimize $L_{s_t} - L_{s_{t^*}}$, which contains the proper setting of the choices for the user during the initial exploration phase, drawing classical user portraits due to the basic messages and give targeted choices, etc.

When the recommending system have collected a certain amount of data, the minimax algorithm, that is, set the weight properly will enhance the effect in the algorithm. The attainment of the mini-max optimal regret is reached without prior insight into the premier arm selection method—including details like the segment count S, the duration of these segments, and their positioning—as well as the duration of the game T. Our findings are universally assertive across every conceivable sequence of arm losses, as they are individually tailored and don't rely on any statistical presuppositions about the behavior of the bandit arms. The proved theorical one gives a certain upper bound of regret, which is algorithm performance.

5.2. Discussions:different recommending systems in the different exploring stage and different individuals

As is mentioned above in the result, flexible design and ultilization of the algorithm in the recommending system may be an advanced choice under the comprehensive consideration on the amount of data, the extent of the typical image of the user, the need of reducing the standard of the regret value and so on. When a recommending system is made, the demand ranking should be clear.

6. Conclusions

In the article, the optimal strategy in the recommending system utilizing the multi-armed bandit is discussed in both aspects: minimizing the total regret and maximize the reward with the limited budget. To reach the optimal reward and minimal comprehensive cost of the exploration phase, the mini-max algorithm is discussed with the combination of the recommending system. It is clear that the algorithms mentioned above is applicable for usage in the recommendation system in many aspects. Supported by a broad range of experiments with real datasets, It can be proved that the mini-max algorithm realizes considerable improvements in performance when bench-marked against leading-edge adversarial multi-armed bandit algorithms found in the literature of reinforcement learning and computational learning theory. Today's existing research achievements concentrate on the regret value's calculating and reducing. With the development of the bandit algorithm and the recommending system, the combination between them has to be closer. For example, the weight in the mini-max algorithm could be explored further; the adjustment of the algorithm strategy during the different stages in the exploration; advancing the arrangement of the cost with the limited budget, etc.

References

[1] Colas, J. T., O'Doherty, J. P., & Grafton, S. T. (2024). Active reinforcement learning versus action bias and hysteresis: Control with a mixture of experts and nonexperts. PLoS Computational Biology, 20(3), Article e1011950. https://doi.org/10.1371/journal.pcbi.1011950

[2] Gokcesu, K., & Kozat, S. S. (2018). An online minimax optimal algorithm for adversarial multiarmed bandit problem. IEEE Transactions on Neural Networks and Learning Systems, 29(11), 5565-5580. https://doi.org/10.1109/TNNLS.2018.2806006

[3] Yang, S., & Gao, Y. (2021). An optimal algorithm for the stochastic bandits while knowing the near-optimal mean reward. IEEE Transactions on Neural Networks and Learning Systems, 32(5), 2285-2291. https://doi.org/10.1109/TNNLS.2020.2995920

[4] Marković, D., Stojić, H., Schwöbel, S., & Kiebel, S. J. (2021). An empirical evaluation of active inference in multi-armed bandits. Neural Networks, 144, 229-246. https://doi.org/10.1016/j.neunet.2021.08.018

[5] Xia, Y., Qin, T., Ding, W., Li, H., Zhang, X., Yu, N., & Liu, T.-Y. (2017). Finite budget analysis of multi-armed bandit problems. Neurocomputing, 258, 13-29.

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

[7] Odeyomi, T. (2020). Learning the truth in social networks using multi-armed bandit. IEEE Access, 8, 137692-137701. https://doi.org/10.1109/ACCESS.2020.3012593

[8] Auer, P., Cesa-Bianchi, N., & Schapire, R. E. (1995). Gambling in a rigged casino: The adversarial multi-armed bandit problem. Proceedings of the 36th Annual Symposium on Foundations of Computer Science, 322-331.

[9] Elena, G., Milos, K., & Eugene, I. (2021). Survey of multiarmed bandit algorithms applied to recommendation systems. International Journal of Open Information Technologies, 9(4), 12-27.

[10] 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.