Proceedings of CONF-MLA 2024 Workshop: Semantic Communication Based Complexity Scalable Image Transmission System for Resource Constrained Devices
DOI: 10.54254/2755-2721/83/2024GLG0074

Comparison of Multi-Armed Bandit Algorithms in Advertising Recommendation Systems

Junyi Zhao

Northwestern Polytechnical University, Xi'an, Shaanxi, 710072, China

zhaojunyi@mail.nwpu.edu.cn

Abstract. In today's rapidly evolving online environment, advertising recommendation systems utilize multi-armed bandit algorithms like dynamic collaborative filtering Thompson sampling (DCTS), upper confidence bound based on recommender system (UCB-RS), and dynamic ε-greedy algorithm (DEG) to optimize ad displays and enhance click-through rates (CTR). These algorithms must adapt to limited information and update strategies based on immediate feedback.This study employs an experimental comparison to assess the performance of the DCTS, UCB-RS, and DEG algorithms using the click-through rate prediction database from Kaggle. Five experimental sets under varied parameter settings were analyzed, employing the Receiver Operating Characteristic (ROC) curve, accuracy, and area under the curve (AUC) metrics.Results show that the DEG algorithm consistently outperforms the others, achieving higher AUC values and demonstrating robust sample identification capabilities. DEG also exhibits superior precision at high recall levels, showcasing its potential in dynamic advertising environments. Its dynamic adjustment strategy effectively balances exploration and exploitation, optimizing ad displays.The findings suggest that DEG's adaptability and stability make it particularly suitable for dynamic ad recommendation scenarios. Future research should focus on optimizing DEG's parameter settings and possibly integrating UCB-RS's exploration mechanisms to enhance performance and develop more effective strategies for advertising recommendation systems.

Keywords: DCTS, UCB-RS, DEG, advertising recommendation systems.

1. Introduction

In today's digital economy, advertising recommendation systems play a crucial role in driving user engagement and revenue, heavily relying on multi-armed bandit algorithms to optimize ad displays and enhance click-through rates (CTR). These algorithms make decisions under limited information and adjust strategies with real-time feedback, thereby improving ad accuracy and return on investment (ROI) by balancing the exploration of new advertising opportunities and the exploitation of known effective ads. Thus, selecting the appropriate algorithm is crucial for effective ad recommendation systems.

The most commonly used algorithms are UCB, TS, and epsilon greedy. Many scholars have improved and perfected the application of these three algorithms. Various authors have applied and improved UCB algorithms through a variety of innovative approaches: Qiu et al. integrated contrastive self-supervised learning with reinforcement learning to propose UCB-type algorithms for MDPs and MGs with low-rank transitions, achieving sample efficiency[1]; He et al. introduced UCBVI-γ, a model-

DOI: 10.54254/2755-2721/83/2024GLG0074

based algorithm for discounted MDPs using optimism and a Bernstein-type bonus, which achieves near-minimax optimal performance[2]; Dong et al. examined the convergence of the MC-UCB algorithm in random-length episodic MDPs, demonstrating almost sure convergence for a large class of MDPs[3]; Lu et al. and Tiapkin et al. introduced Causal MDPs and the C-UCBVI algorithm, leveraging causal structures to improve reinforcement learning performance and achieve efficient regret bounds[4, 5]; Domingues et al. developed Kernel-UCBVI, an optimistic algorithm for finite-horizon RL problems with a metric state-action space, using kernel estimators to balance exploration and exploitation and achieve novel regret bounds[6]; and Foster and Rakhlin proposed a universal and optimal reduction from contextual bandits to online regression, utilizing regression oracles and UCB-based exploration strategies to achieve minimax optimal rates for general function classes[7].

Many scholars have explored and advanced the applications of Thompson Sampling (TS) algorithms in different domains. Zhang et al. introduced Neural Thompson Sampling, utilizing deep neural networks to balance exploration and exploitation in contextual multi-armed bandit problems, achieving a cumulative regret of O(T1/2) and demonstrating strong experimental performance[8]. Aouali et al. developed Mixed-Effect Thompson Sampling (meTS) for contextual bandits, using a mixed-effect model to capture action correlations and providing bounds on Bayes regret with strong empirical results in both synthetic and real-world scenarios[9]. Peng and Zhang demonstrated that top-two Thompson Sampling excels in ranking and selection problems, offering comprehensive theoretical and numerical comparisons with other sampling procedures[10]. Uguina et al. introduced a learnheuristic algorithm that combines Thompson Sampling with metaheuristics to solve the dynamic team orienteering problem, adapting to real-time changes such as traffic and weather, and outperforming static approaches by up to 25% in dynamic settings[11]. Tanık and Ertekin proposed a hierarchical framework that integrates reinforcement learning and a TS-inspired soft-attention model to improve learning and planning through skill reuse and adaptability, efficiently solving compositional control problems[12]. Finally, Bi et al. enhanced personalized dynamic pricing with an improved Thompson Sampling algorithm, leveraging the Pólya-Gamma distribution to handle high-dimensional feature vectors, resulting in faster convergence and lower regret compared to traditional methods in both simulated and real data[13].

The epsilon-greedy algorithm has been advanced by various researchers through diverse applications and enhancements. You et al. introduced the EMMA algorithm, which leverages epsilon-greedy for optimizing MQTT QoS mode selection and power control in power distribution IoT (PD-IoT), balancing packet-loss ratio and energy consumption via online learning in a multi-armed bandit framework[14]. Yang et al. developed an adaptive epsilon-greedy strategy for a multi-objective hyper-heuristic algorithm (HH_EG) that effectively selects and combines low-level heuristics (LLHs) during evolution, enhancing cross-domain problem-solving without redesigning high-level strategies[15]. Liu et al. enhanced particle swarm optimization (PSO) with an epsilon-greedy strategy and a Pareto archive algorithm for multi-objective reactive power optimization, improving global and local search capabilities to explore optimal solutions early and avoid local optima[16]. Dabney et al. proposed a temporally extended epsilon-greedy algorithm that improves exploration by repeating actions for random durations, inspired by ecological models, showing strong performance under certain duration distributions[17]. Gimelfarb et al. introduced a Bayesian ensemble approach (ε-BMC) to epsilon-greedy exploration in model-free reinforcement learning, with a closed-form Bayesian model update to adapt ε efficiently, balancing exploration and exploitation with monotone convergence guarantees[18]. Finally, Rawson and Balan developed the Deep Epsilon Greedy method, using neural network predictions to achieve error or regret bounds and convergence guarantees, demonstrating that cubic root exploration minimizes the regret upper bound in nonlinear reinforcement learning problems, as evidenced by experiments on the MNIST dataset[19].

Despite the advancements in multi-armed bandit algorithms, there remains a significant gap in their application to advertising recommendation systems operating in dynamic, non-static environments. Traditional algorithms are often evaluated under static conditions, which do not adequately reflect the rapid changes in user preferences and behaviors encountered in real-world advertising scenarios. Through a detailed comparative analysis of DCTS, UCB-RS, and DEG, this study aims to provide

DOI: 10.54254/2755-2721/83/2024GLG0074

valuable insights into their performance under real-time dynamic conditions, ultimately guiding practitioners in selecting the most effective algorithms for advertising recommendation systems.

The effects of each algorithm are verified in combination with experimental data sets, especially in terms of accuracy and efficiency of advertising recommendation, providing an empirical basis for algorithm selection and parameter adjustment of advertising systems. The comparison and optimization of these algorithms not only improves the relevance and user satisfaction of advertising recommendations, but also provides a more flexible and adaptable solution for the field of advertising recommendation.

2. Methodology

2.1. Thompson sapling

The application of Thompson sapling(TS) in ad recommendation mainly relies on its ability to balance exploration and utilization to maximize long-term benefits.

TS models the performance of each ad, usually using Beta distribution to represent the uncertainty of the success rate (click-through rate) of each ad. Each ad is assigned two parameters, the number of successes (α) and the number of failures (β). Each time an ad is displayed, the TS algorithm extracts a probability value for each ad, which comes from the Beta distribution of the ad success rate. Then the ad with the highest probability value is selected to be displayed to the user. This method naturally balances the utilization of known efficient ads and the exploration of uncertain ads. Whenever an ad is clicked or ignored, the Beta distribution parameters (α or β) of the corresponding ad are updated to reflect the latest user feedback. This instant update allows the model to quickly adapt to changes in user behavior. By continuously optimizing these parameters, TS can improve the adaptability of the ad recommendation system to user preferences, thereby increasing the user's click-through rate and conversion rate, and ultimately achieve the goal of maximizing advertising revenue[20].

This paper uses a model called DCTS for comparison, which is designed for cross-domain ad recommendations. This model leverages the inherent similarity between users and ads and enhances the traditional Thompson sampling method by combining temporal dynamics and cross-domain knowledge transfer. By more accurately predicting user preferences over time and in different ad environments, DCTS significantly improves CTR[21].

When updating the Beta distribution parameters, DCTS considers the information from the global reward and the feedback from individual users. The parameter update formula is as follows:

Posterior distribution parameters:

$$\alpha_k(t) = \lambda(s)\alpha_{0,k}(t) + gs_k(t) \tag{1}$$

$$\beta_k(t) = \lambda(f)\beta_{0,k}(t) + gf_k(t) + 1 \tag{2}$$

Where,$\lambda(s)$ and $\lambda(f)$ are hyperparameters that adjusts the importance of prior knowledge.They scale the impact of initial or baseline parameters ($\alpha_{0,k}(t)$) and ($\beta_{0,k}(t)$) on the current calculation of the Beta distribution's parameters.g is a hyperparameter that adjusts the importance of global reward, and $s_k(t)$ and $f_k(t)$ are the number of successes and failures at time step t, respectively[21].

2.2. Upper Confidence Bound

The Upper Confidence Bound (UCB) algorithm is a powerful reinforcement learning technique used in market recommendation systems, particularly for optimizing ad placements. In dynamic environments where customer preferences evolve over time, UCB effectively balances exploration (trying new ads) and exploitation (favoring the best-performing ads). It does so by estimating the potential reward of each ad and adjusting the selection strategy accordingly. This approach ensures that the system not only leverages historical data but also adapts to changes, maximizing long-term CTR while reducing the risk of missing out on better-performing ads[22].

DOI: 10.54254/2755-2721/83/2024GLG0074

This paper adopts the improved algorithm UCB-RS (Recommendation system-based Upper Confidence Bound) of the UCB algorithm[23]. This algorithm performs well in dealing with multi-armed bandit problems with non-stationary and large-scale state spaces by combining the advantages of upper confidence bounds and recommendation systems.

The advertising recommendation system needs to choose from a large number of possible advertisements, which makes the pure exploration phase very time-consuming. The UCB-RS algorithm greatly reduces the time of pure exploration through the collaborative filtering technology of the recommendation system. Specifically, user-user collaborative filtering is used to estimate the potential rewards of unselected advertisements, so as to directly enter the utilization phase and improve the efficiency of the algorithm.

Main Formulas and Parameters

1) Estimated Mean Reward:

$$\mu_{it} = \lambda\mu_{it} + (1-\lambda)\widehat{\mu_{it}} \tag{3}$$

Where λ is the coefficient that balances the real-time reward (μ_{it}) and the reward estimated by the recommendation system ($\widehat{\mu_{it}}$), ranging between 0 and 1.

2) Upper Confidence Bound):

$$U_{kt} = \mu_{it} + \xi_{it} \tag{4}$$

Where ξ_{it} is the confidence interval, defined as:

if $N_i = 0$:

$$\xi_{it} = \sqrt{\alpha \log T} \tag{5}$$

otherwise:

$$\xi_{it} = \sqrt{\alpha \log \frac{T}{N_i}} \tag{6}$$

N_i is the number of times ad i has been played up to time t, α is the confidence level parameter, T is a predefined time interval much longer than most ad sessions.

3) Recommendation System Framework:

Uses Collaborative Filtering (CF) to compute the potential rewards for each product based on the historical data of similar users in the reference set.

Parameter Configuration

- λ: Coefficient balancing real-time and estimated rewards, ranges from 0 to 1.
- α: Confidence level parameter, affects the size of the confidence interval.
- T: Predefined long time interval, used for calculating the confidence interval of unexplored ads.
- N_i: Number of times ad i has been played up to time t[23].

2.3. Epsilon Greedy

The ε-greedy algorithm is an algorithm used to solve the multi-armed bandit problem. The algorithm dynamically adjusts to obtain the maximum benefit in the long term by exploring and exploiting with a certain probability. The specific approach is: explore with probability ε, select random selection options to discover potential high-yield options; exploit with probability 1-ε, select the option with the highest estimated benefit at present. This algorithm can explore different options while giving priority to the current best benefit option to obtain the maximum benefit.

The ε-greedy algorithm is applied to the advertising recommendation system. The algorithm can attract customers by showing them different content, thereby obtaining higher CTR and conversion rates. By adjusting ε in the algorithm, the advertising recommendation system can find a balance between exploring new ads and utilizing the existing best-performing ads, thereby maximizing benefits.

DOI: 10.54254/2755-2721/83/2024GLG0074

The core of using the ε-greedy algorithm is to adjust the ε value. Too high or too low ε will lead to poor performance of the algorithm. The actual performance is: at a higher ε value, the system may explore too frequently, resulting in the selection of many suboptimal discount strategies, which ultimately reduces the overall benefit. At a lower ε value, the system may rely too much on existing efficient strategies and fail to discover potential higher-yield options, which will also lead to suboptimal performance. This paper adopts the ε-greedy algorithm that dynamically adjusts the ε value to optimize the decision-making of data based on the user's behavioral preferences. Specifically, this algorithm sets a higher ε value at the beginning and conducts extensive exploration; and gradually reduces the ε value over time to increase the utilization of the current optimal option[24].

This paper selected the dynamic adjustment of the Epsilon Greedy algorithm for application. Because it is crucial for effectively balancing exploration with exploitation, especially in environments where options or states are numerous and varied[24]. Below is the comprehensive formula and the step-by-step application process:

Formula

Epsilon (ε) Adjustment Formula:

$$\varepsilon(t) = \max\left(\varepsilon_{\min}, \varepsilon_0 * e^{-r*t}\right) \tag{7}$$

Where ε_0 is the initial value of epsilon, r is the decay rate, t is the iteration or time step, and ε_{min} is the minimum value of epsilon to ensure some degree of exploration continues throughout the learning process.

Application Process

1)Initialize Parameters: Set ε_0, r, ε_{min}, and initialize the decision environment and reward structure.

2)For each iteration t:

- Calculate ε(t) using the formula based on the current iteration number.
- Decision Selection:
  - With probability ε(t), select a random action (explore).
  - With probability 1 - ε(t), select the best-known action (exploit).
- Execute the action and receive a reward.
- Update reward estimation for the chosen action based on the received reward.
- Periodically evaluate or monitor algorithm performance to determine if adjustments to decay rate or other parameters are needed.

3)Termination Condition:

- Reach a predetermined number of iterations.
- Performance reaches a satisfactory level or no significant improvement is observed.

3. Research design

3.1. Objective

This study aims to evaluate the performance of three multi-armed bandit algorithms (DCTS, UCB-RS, and DEG) in advertising recommendation systems. The specific goal is to compare the performance of these algorithms under different parameter settings through experiments, so as to provide optimization strategies for advertising recommendation systems in dynamic non-static environments.

3.2. Experimental Groups

In order to comprehensively evaluate the performance of the algorithms under different conditions, this study designed five experimental groups, each with different parameter configurations to simulate different advertising market environments and user behaviors, which are shown in Table1. The settings of the parameters reflect the balance between historical data and new data, and the trade-off between exploration and utilization. Specifically, in the DCTS algorithm, when the Lambda value is higher (such as 0.8), the algorithm relies more on historical data, while a lower Lambda value (such as 0.65) increases the tendency to explore new ads. The UCB-RS algorithm controls the intensity of exploration by

DOI: 10.54254/2755-2721/83/2024GLG0074

adjusting the values of Lambda and Alpha. In scenarios with high uncertainty in ad click-through rates, higher Lambda and Alpha values (such as 0.6 and 2.5) can prompt the algorithm to explore new ads more, while lower values (such as 0.4 and 1.8) make the algorithm more inclined to use existing ad recommendations. The Epsilon0 and Decay Rate parameter settings in the DEG algorithm affect the exploration degree and convergence speed of the algorithm in the initial stage. A higher Epsilon0 and a lower Decay Rate (such as 0.35 and 0.0008) allow the algorithm to maintain a higher exploration rate for a longer period of time, which helps to discover new efficient advertisements, while a lower Epsilon0 and a higher Decay Rate (such as 0.25 and 0.0012) are suitable for a relatively stable advertising market. These parameter configurations balance exploration and utilization, historical data and new data, and are suitable for advertising recommendation systems under different uncertainties and market dynamics.

Table 1. Parameters settings

Experimental Group	DCTS Lambda	DCTS g	DCTS Gamma	UCB-RS Lambda	UCB-RS Alpha	DEG Epsilon0	DEG Decay Rate
Group 1	0.75	0.3	0.25	0.5	2	0.3	0.001
Group 2	0.85	0.4	0.2	0.6	2.5	0.35	0.0008
Group 3	0.65	0.2	0.3	0.4	1.8	0.25	0.0012
Group 4	0.7	0.25	0.35	0.45	2.2	0.28	0.001
Group 5	0.8	0.35	0.22	0.55	2.1	0.32	0.0009

3.3. Performance Metrics

This study uses ROC curve, AUC value and average precision (AP) as the main performance evaluation indicators. ROC curve is used to analyze the performance of the algorithm under different thresholds, and the algorithm's ability to distinguish is judged by observing the shape and position of the curve. The AUC value, as the area under the ROC curve, directly quantifies the overall ability of the algorithm to distinguish positive and negative samples. The higher the AUC value, the better the algorithm performance. In addition, the average precision (AP) is used to evaluate the accuracy of the algorithm under high recall rate, which is particularly important for advertising recommendation systems because in real applications, high recall rate usually means higher user coverage and advertising click-through rate.

3.4. Experimental Procedure

3.4.1. Data Preprocessing

The experiment first preprocesses the click-through rate prediction database from the Kaggle website, including deleting abnormal data groups to ensure the reliability of the experimental data. Then, for the missing values in the data, the average value of the overall data is used to fill in the missing values to ensure the stability and repeatability of the experimental results.

3.4.2. Parameter Configuration

Different algorithm parameters are set for each experimental group to test their impact on the performance of the advertising recommendation system. Specifically, the DCTS algorithm adjusts the Lambda, g, and Gamma values to study its adaptability to different advertising environments; the UCB-RS algorithm optimizes its performance in uncertain markets by changing the Lambda and Alpha values; the DEG algorithm controls the balance between exploration and utilization by adjusting Epsilon0 and Decay Rate.

DOI: 10.54254/2755-2721/83/2024GLG0074

3.4.3. Algorithm Execution

In each experimental group, the DCTS, UCB-RS, and DEG algorithms are executed to predict and optimize the click-through rate by simulating the actual scenario in the advertising recommendation system to test the performance of each algorithm under different parameter settings.

3.4.4. Data Analysis

Finally, by comparing and analyzing the ROC curves, AUC values, and average precision of each experimental group, the impact of different parameter settings on the algorithm performance is discussed in detail, and then the optimal algorithm configuration scheme under different advertising market conditions is determined.

4. Research results

It can be seen from Table2 that in all experiments, the DEG algorithm has the highest AUC value, which is always between 0.78 and 0.79, indicating that it performs best in distinguishing positive and negative samples. UCB-RS is second, with an AUC value of about 0.77. DCTS has the lowest AUC value, mainly between 0.74 and 0.76; at the same time, the DEG algorithm always shows the highest average precision (AP) in all experiments, mainly between 0.77 and 0.78. The average precision of UCB-RS is second, concentrated between 0.74 and 0.75. DCTS has the lowest average precision, between 0.73 and 0.76.

Table 2. Research result

Experiment	DCTS AUC	DCTS AP	UCB-RS AUC	UCB-RS AP	DEG AUC	DEG AP
1	0.75	0.73	0.77	0.75	0.79	0.77
2	0.74	0.72	0.77	0.74	0.78	0.77
3	0.76	0.76	0.77	0.75	0.79	0.78
4	0.75	0.74	0.77	0.74	0.79	0.78
5	0.74	0.72	0.77	0.74	0.79	0.77

By analyzing the ROC curves of the three algorithms from Figure1 and Figure2, it can be seen that the curve trends of the three algorithms are very close, and the AUC values are also relatively similar. This shows that under these experimental settings, the performance of the three algorithms in distinguishing positive and negative samples is not much different. In most experiments, the AUC value of the DEG algorithm is slightly higher than that of the other two algorithms, which means that under the given experimental settings, the overall performance of the DEG algorithm under the ROC curve is slightly better.

DOI: 10.54254/2755-2721/83/2024GLG0074

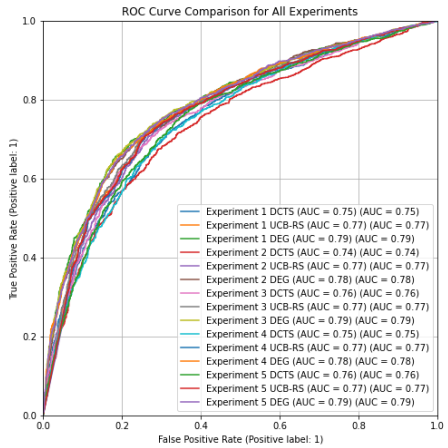

Figure 1. ROC Curve Comparison for All Experiments

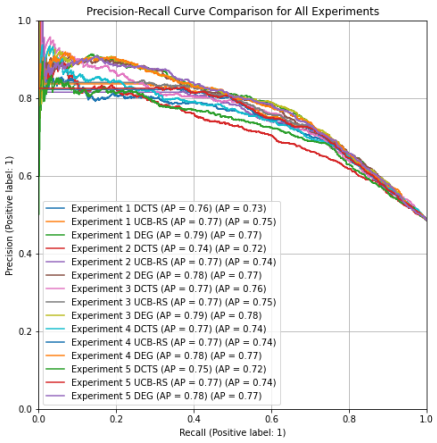

Figure 2. Precision-Recall Curve Comparison for All Experiment

In general, in all experiments, DEG has the highest AUC and AP values, indicating that it can not only effectively distinguish positive and negative samples in advertising recommendation, but also maintain a high precision at a high recall rate. For UCB-RS, although it does not perform as well as DEG, it is still stable in AUC and AP values, indicating that it has good balance and robustness in advertising recommendation. In all experiments, DCTS has the lowest AUC and AP values, especially at high recall rates, where its precision drops rapidly, indicating that it has a weaker ability to adapt to new ads.

5. Result analysis

After comparatively analyzing the Precision-Recall curves of the three algorithms, we found that the Dynamic Epsilon Greedy (DEG) algorithm can maintain high precision even under high recall, which shows that it has significant advantages in identifying positive samples. The performance of the UCB-RS and DCTS algorithms is relatively stable under high recall conditions, although their accuracy has declined in some tests, especially DCTS, which performed poorly in some tests. Taken together, the DEG algorithm shows the best AUC and accuracy in multiple experiments. The DEG algorithm successfully balances the click-through rate and advertising display effect by dynamically adjusting the exploration and utilization strategies, making it the best-performing algorithm.

The main reason why the DEG algorithm performs well is that it dynamically balances the relationship between exploration and utilization. Especially in the field of advertising recommendation, the high degree of uncertainty and variability in this field places higher requirements on the adaptability and robustness of the algorithm. The DEG algorithm continuously adjusts the epsilon value and dynamically adjusts the proportion of exploration and utilization according to environmental changes, so that it can still maintain high performance when facing different advertising recommendation scenarios. In contrast, although the UCB-RS algorithm performs well in some scenarios, its exploration mechanism is relatively fixed and cannot respond to different market environments as flexibly as the DEG algorithm. The poor performance of the DCTS algorithm may be attributed to its insufficient precision when dealing with high recall rates, which indicates that it has certain limitations in the identification of positive and negative samples.

Based on the experimental results, it is recommended to prioritize the DEG algorithm in actual advertising recommendation systems and further optimize its parameters to improve performance. The DEG algorithm can not only effectively respond to changes in the advertising market, but also achieve the best balance between click-through rate and display effect. Although DEG performs well, in future work, it can be considered to combine the exploration mechanism of the UCB-RS algorithm with the

DOI: 10.54254/2755-2721/83/2024GLG0074

DEG algorithm to further improve its performance in the dynamic advertising market. In addition, real-time adjustment strategies as the market changes are also worth considering, which will help improve the effectiveness of the advertising recommendation system in various scenarios.

The DCTS algorithm leverages historical medical data to recommend treatments for chronic diseases, making it effective in data-rich scenarios but less so with new or rare conditions due to its reliance on past data. The UCB-RS algorithm balances the exploration of new treatments with the utilization of known ones by calculating confidence intervals, suitable for complex medical scenarios where treatment efficacy is uncertain. The DEG algorithm adapts dynamically between exploring and exploiting treatment options, ideal for personalized medicine and chronic disease management, although it requires rigorous risk management to ensure safety and effectiveness in its exploratory approaches.

6. Conclusion

This study evaluates the performance of three multi-armed bandit algorithms—DCTS, UCB-RS, and DEG—in advertising recommendation systems. Comparative analysis found that in dynamic environments, the DEG algorithm performed superiorly in terms of AUC. The DEG algorithm can effectively balance exploration and utilization, optimize advertising display based on real-time data, and demonstrate excellent adaptability under rapidly changing user preferences. These findings suggest that advertising recommendation systems should prioritize the implementation of flexible algorithms like DEG that can swiftly adapt to changing user behaviors, thereby enhancing user engagement and CTR.

The research results show that the DEG algorithm maintained the highest AUC value in all experiments, indicating that it has significant advantages in distinguishing positive and negative samples. In addition, the DEG algorithm can still maintain high precision under high recall, which emphasizes its application potential in dynamic advertising recommendation. These conclusions not only provide scientific basis for algorithm selection and parameter adjustment of advertising recommendation systems, but also point out the direction for the development of advertising recommendation technology.

In terms of implications for future research, the findings of this paper help close the knowledge gap in applying multi-armed bandit algorithms in non-static environments. Future research should focus on optimizing the parameter settings of the DEG algorithm and exploring hybrid approaches that integrate the exploration mechanisms of UCB-RS, aiming to enhance performance in dynamic advertising contexts. In addition, the study also suggests combining the exploration mechanism of UCB-RS with DEG to further improve the performance of the algorithm in the dynamic advertising market. This study not only contributes to the theoretical understanding of multi-armed bandit algorithms in advertising but also offers practical insights for their implementation, paving the way for future research in adaptive recommendation systems.

References

[1] Qiu, Shuang, et al. "Contrastive ucb: Provably efficient contrastive self-supervised learning in online reinforcement learning." International Conference on Machine Learning. PMLR, 2022.

[2] He, Jiafan, Dongruo Zhou, and Quanquan Gu. "Nearly minimax optimal reinforcement learning for discounted MDPs." Advances in Neural Information Processing Systems 34 (2021): 22288-22300.

[3] Dong, Zixuan, Che Wang and Keith W. Ross. “On the Convergence of Monte Carlo UCB for Random-Length Episodic MDPs.” ArXiv abs/2209.02864 (2022): n. pag.

[4] Lu, Yangyi, Amirhossein Meisami and Ambuj Tewari. “Causal Markov Decision Processes: Learning Good Interventions Efficiently.” ArXiv abs/2102.07663 (2021): n. pag.

[5] Daniil Tiapkin, Denis Belomestny, Eric Moulines, Alexey Naumov, Sergey Samsonov, Yunhao Tang, Michal Valko, Pierre Menard. “From Dirichlet to Rubin: Optimistic Exploration in RL without Bonuses.” The 39th International Conference on Machine Learning, PMLR 162:21380-21431, 2022.

DOI: 10.54254/2755-2721/83/2024GLG0074

[6] Omar Darwiche Domingues, Pierre Menard, Matteo Pirotta, Emilie Kaufmann, Michal Valko. “Kernel-Based Reinforcement Learning: A Finite-Time Analysis.” The 38th International Conference on Machine Learning, PMLR 139:2783-2792, 2021.
[7] Dylan Foster, Alexander Rakhlin. “Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles.” The 37th International Conference on Machine Learning, PMLR 119:3199-3210, 2020.
[8] Zhang, W., Zhou, D., Li, L., & Gu, Q. (2020). Neural thompson sampling. arXiv preprint arXiv:2010.00827.
[9] Aouali, I., Kveton, B., & Katariya, S. (2023, April). Mixed-effect thompson sampling. In International Conference on Artificial Intelligence and Statistics (pp. 2087-2115). PMLR.
[10] Peng, Y., & Zhang, G. (2022, December). Thompson sampling meets ranking and selection. In 2022 Winter Simulation Conference (WSC) (pp. 3075-3086). IEEE.
[11] Uguina, A. R., Gomez, J. F., Panadero, J., Martínez-Gavara, A., & Juan, A. A. (2024). A Learnheuristic Algorithm Based on Thompson Sampling for the Heterogeneous and Dynamic Team Orienteering Problem. Mathematics, 12(11), 1758.
[12] Tanık, Güven Orkun, and Şeyda Ertekin. "Hierarchical reinforcement Thompson composition." Neural Computing and Applications (2024): 1-10.
[13] Bi, Wenjie, Bing Wang, and Haiying Liu. "Personalized Dynamic Pricing Based on Improved Thompson Sampling." Mathematics 12.8 (2024): 1123.
[14] You, X., Zhang, P., Liu, M., Lin, L., & Li, S. (2023). Epsilon-Greedy-Based MQTT QoS Mode Selection and Power Control Algorithm for Power Distribution IoT. International Journal of Mobile Computing and Multimedia Communications (IJMCMC), 14(1), 1-18.
[15] Yang, T., Zhang, S., & Li, C. (2021). A multi-objective hyper-heuristic algorithm based on adaptive epsilon-greedy selection. Complex & Intelligent Systems, 7, 765-780.
[16] Liu, X., Zhang, P., Fang, H., & Zhou, Y. (2021). Multi-objective reactive power optimization based on improved particle swarm optimization with ε-greedy strategy and pareto archive algorithm. IEEE Access, 9, 65650-65659.
[17] Dabney, W., Ostrovski, G., & Barreto, A. (2020). Temporally-extended {\epsilon}-greedy exploration. arXiv preprint arXiv:2006.01782.
[18] Gimelfarb, M., Sanner, S., & Lee, C. G. (2020). {\epsilon}-bmc: A bayesian ensemble approach to epsilon-greedy exploration in model-free reinforcement learning. arXiv preprint arXiv:2007.00869.
[19] Rawson, M., & Balan, R. (2021). Convergence guarantees for deep epsilon greedy policy learning. arXiv preprint arXiv:2112.03376.
[20] Pant, K. A., Hegde, A., & Srinivas, K. V. (2022). Thompson Sampling with Virtual Helping Agents. arXiv preprint arXiv:2209.08197.
[21] Ishikawa, S., Chung, Y. J., & Hirate, Y. (2022). Dynamic collaborative filtering Thompson Sampling for cross-domain advertisements recommendation. arXiv preprint arXiv:2208.11926.
[22] Singh, V., Nanavati, B., Kar, A. K., & Gupta, A. (2023). How to maximize clicks for display advertisement in digital marketing? A reinforcement learning approach. Information Systems Frontiers, 25(4), 1621-1638.
[23] Nguyen-Thanh, N., Marinca, D., Khawam, K., Rohde, D., Vasile, F., Lohan, E. S., ... & Quadri, D. (2019). Recommendation system-based upper confidence bound for online advertising. arXiv preprint arXiv:1909.04190.
[24] Ishikawa, S., Chung, Y. J., & Hirate, Y. (2022). Dynamic collaborative filtering Thompson Sampling for cross-domain advertisements recommendation. arXiv preprint arXiv:2208.11926.