Clustered UCB-C: Exploiting User Heterogeneity for Low-Regret Recommendation in Structured Multi-Armed Bandit Settings

Zhitao Wang

{w18618118890@163.com}

The Experimental High School Affiliated To Beijing Normal University, Beijing, 100032, China

Abstract. The heterogeneity of user preferences is a core challenge in recommendation systems. Traditional MAB algorithms have limitations such as single clustering and fixed α. Therefore, it is crucial to enhance their practicality in recommendation systems. Based on the MovieLens 1M dataset, this paper proposes a two-step strategy of "preference clustering and users clustering" and personalized α allocation, and designs the clustering UCB-C algorithm to study low regret recommendations. Experiments revealed 8 user groups (covering 3 types, such as Drama), and the cumulative regret of this algorithm was significantly lower than that of the ordinary UCB, with the regret growth slowing down in the middle and high rounds. The personalized α adapts to the exploration needs of different groups. This research provides a reusable framework for heterogeneous scenario recommendations and has significant practical value.

Keywords: UCB-C algorithms, heterogeneity, cumulative regret, recommendation systems.

1 Introduction

In the age of rapid development of the Internet, recommendation systems have become a crucial tool for processing data and achieving personalized recommendations. It is widely applied in fields such as film and television, short videos, and e-commerce. The preference differences among heterogeneous user groups (such as the varying demands for different types of movies among groups of different ages or occupations) are the core challenge faced by recommendation systems. The balance issue between exploring unknown preferences and utilizing known preferences directly affects the effectiveness of recommendations. The Multi-Armed Bandit (MAB) algorithm, with its machine learning feature, provides an effective solution to this problem. Among them, the Upper Confidence Bound (UCB) series of algorithms, due to their rigorous theory, are widely applied in recommendation scenarios [1][2].

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365117

The current research has made certain progress: Mukherjee et al. proposed a unified framework for structured MAB, designed the UCB-C algorithm to extend the classic Bandit to the structured scenario, and proved that non-competitive arms only need to be pulled O(1) times to reduce regret [3][4]. Gore et al. proposed the Clus-UCB algorithm, which constructs a clustering adaptation mechanism through the KL divergence confidence interval, and the derived asymptotic regret lower bound is superior to traditional MAB algorithms [5]. Ban et al. proposed the Local Clustering in Bandits (LOCB) algorithm, which realizes local user clustering in contextual Bandit based on incremental estimation of unknown parameters, and adapts to heterogeneous scenarios[6].

However, these methods have limitations: Firstly, traditional structured MAB does not adequately correlate user attributes with preferences, and the clustering strategies are mostly divided along a single dimension, making it difficult to adapt to complex heterogeneous scenarios. The LOCB algorithm, although it achieves local clustering, does not consider the processing requirements for mixed data types. Secondly, the exploration coefficient α of the UCB algorithm is mostly a globally fixed value. This cannot match the exploration needs of different user groups. Thirdly, some algorithms have not been sufficiently validated on real datasets, and the computational resource consumption is relatively high.

This paper focuses on the application of the Clustered UCB-C Algorithm in heterogeneous recommendation systems. This paper proposes a two-step grouping strategy of "preference grouping and attribute segmentation", and designs a personalized α allocation mechanism based on group heterogeneity. This provides a solution to the problem that the traditional UCB algorithm cannot meet the preference differences among different user groups in the application of recommendation systems. Through multi-dimensional experiments on the MovieLens dataset, a reusable experimental framework is provided to verify the effectiveness of the algorithm.

The subsequent sections of the paper are arranged as follows: The second section introduces the theoretical basis and provides a detailed description of the methods used in this study; the third section analyzes the experimental results; the fourth section discusses the limitations of the algorithm and proposes possible solutions.

2 Methodology

2.1 Data Introduction

Harper and Konstan proposed the MovieLens 1M dataset, which contains 1 million ratings from 6,040 users. The richness of user attributes and movie features in this dataset highlights the difficulty in recommendations due to user heterogeneity [7]. Therefore, designing a clustering UCB algorithm that is suitable for this scenario can reduce cumulative regret and has significant theoretical and practical value for improving the practicality of recommendation systems.

2.2 Problem Formulation

The basic model of the multi-armed bandit. In the MAB problem, suppose the recommendation system has K candidate items (arms), and the reward of each arm i follows an independent and identically distributed $X_i \sim F_i(\mu_i)$, where μ_i is the unknown mean. The algorithm selects an arm a_t to pull in each round t observes the reward x_t, and updates the strategy. The goal is to minimize the cumulative regret:

$$Reg(T) \;=\; E\;[\Sigma_{t=1}^{T}(\mu^{*}-\mu_{a_t})] \tag{1}$$

where μ^* is the mean of the optimal arm [8].

The core formula of the UCB algorithm is:

$$UCB_i(t)=\hat{\mu_i}(t)+\sqrt{\frac{\alpha logT}{n_i(t)}} \tag{2}$$

here $\hat{\mu_i}(t)$ represents the empirical mean, $n_i(t)$ is the number of pulls, and α is the exploration coefficient. The parameter sensitivity study shows that a smaller α value tends to exploit, while a larger α value tends to explore [8][9].

Symbol Definitions. Based on the multi-armed bandit (MAB) theoretical framework and the characteristics of heterogeneous recommendation scenarios, the core symbols are defined as follows:

Arm set (movie set) $A=\{a_1,a_2,\ldots,a_K\}$, where $K\;=10$ (Retain the ten main movie genres);

Group set $G=\{g_1,g_2,\ldots,g_8\}$through two-step clustering, finally retain 8 user groups. Reward

Reward Assumptions. Based on the characteristics and structure of the MovieLens dataset and the theory of Structured MAB, two hypotheses are proposed:

Core assumption 1: Users in the same group have similar preferences for movie genres, while users in different groups have significantly different preferences.

Core assumption 2: Users' age, occupation, and movie genre preferences are strongly related [3].

Definition of Cumulative Regret. The algorithm aims to minimize the cumulative regret after T rounds. The regret definitions for the clustered UCB-C and the ordinary UCB are distinguished as follows:

Clustered UCB-C: Based on the optimal reward within the cluster, it is defined as

$$\mathrm{Reg(T)}=\mathrm{E}[\sum_{t=1}^{T}(\mu^{*}_{Gt}-\mu_{Gt,at})] \tag{3}$$

where $^{*}_{Gt}$ is the average optimal arm value within cluster G_t in the round t, and at is the recommended movie in that round.

Ordinary UCB: Based on the global optimal reward, it is defined as $\mathrm{Reg(T)} = \mathrm{E}\;[\Sigma_{t=1}^{T}(\mu^{*}-\mu_{a_t})]$, where $\mu*$ is the global optimal arm average value. It serves as a comparison benchmark [8].

2.3 Clustered UCB-C Algorithm Framework

Step 1: Data Preprocessing. Based on the characteristics of the MovieLens 1M dataset, this paper mainly carried out preprocessing operations such as dataset filtering and feature extraction. Firstly, movies were selected. This paper retained movies with clear main genres (such as drama, suspense, crime, comedy, action, etc.); then, users were filtered out. Users with fewer than 20 rating instances were excluded; next, the age and occupation of the selected users were extracted. Finally, the average rating of users for each movie genre was calculated, as well as the preference concentration (defined as the reciprocal of the square root of the variance of type ratings, $1/\sigma^2$), to quantify the distribution of user preferences [7].

Step 2: Two-Stage User Clustering. First, users are grouped based on their preferences for movies. Using the K-Means algorithm, clustering is performed based on the general ratings of movie types and the concentration of preferences of users. A fixed number of clusters of 4 is set, resulting in 4 preference groups. Then, within each preference group, users' age and occupation characteristics are extracted, and the K-Means algorithm is used to complete the subgroup division, resulting in a total of 8 user groups.

Step 3: Competitive Arm Selection. Based on the distribution characteristics of the group's ratings, select the competitive arms to reduce ineffective exploration [10]. First, calculate the average rating $\bar{r}_{g,c}$ and the standard deviation$\sigma_{g,c}$for each movie type within the current group; then, set the threshold as $\max(0.2, \sigma_{g,best} \times 0.8)$, where $\sigma_{g,best}$ is the standard deviation of the optimal type within the group. Finally, retain the movie types that satisfy$\bar{r}_{g,c} \geq \bar{r}_{g,best} -$ threshold as the competing arms of this group [3].

Step 4: Personalized Exploration Strength Allocation. Based on the stability of group preferences and the age characteristics of users, a personalized α allocation mechanism is designed:

High exploration intensity ($\alpha = 5.0$): When σ_g^2 (Group preference variance) < 0.04 and age_g (Group average age) < 30, the young user group has dispersed preferences, and it is necessary to enhance the exploration of unknown preferences;

Medium exploration intensity ($\alpha = 3.0$): When $0.04 \leq \sigma_g^2 < 0.07$, the preferences are moderately dispersed, and a balance between exploration and exploitation should be maintained;

Low exploration intensity ($\alpha = 1.0$): When $\sigma_g^2 \geq 0.07$ and $age_g \geq 45$, the older user group has stable preferences, and emphasis is placed on utilizing known preferences [2].

Step 5: UCB-C Execution. According to the following formula:

$$UCB_{g,a}(t) = \hat{\mu}_{g,a}(t) + \sqrt{\frac{\alpha_g \log T}{n_{g,a}(t)}} \tag{4}$$

Calculate the UCB value for all competing arms within group g.. Here:

$\hat{\mu}_{g,a}(t)$ is the cumulative average reward of group g for the competing arm a in the round t;

$n_{g,a}(t)$ is the number of times the competing arm a in group g was selected before the round t; α_g is the personalized exploration coefficient of group g.; T is the total number of experiment rounds, and t is the current recommendation round.

Then, the movie genre corresponding to the competitive arm with the largest UCB value is selected for recommendation.

Next, receive the user rating feedback and update the cumulative average reward $\hat{\mu}_{g,a}(t)$ and the number of selections $n_{g,a}(t)$ for this competitive arm in the group.

Finally, calculate the regret value: update the cumulative regret in real time and complete one iteration.

3 Result Analysis

3.1 User groups data analysis

Table 1. Main information of the eight groups

	Average age	Main occupation	Optimal type	Optimal score	Sample size
Group1	29.1	other	Drama	3.89	247
Group2	30.3	manager	Drama	3.84	305
Group3	29.9	other	Drama	3.98	273
Group4	24.9	profession	Crime	4.01	290
Group5	29.7	profession	Mystery	4.09	349
Group6	24.5	manager	Crime	4.10	613
Group7	50.4	manger	Mystery	3.44	572
Group8	49.6	other	Mystery	3.44	371

From the group information as shown in Table 1, it can be seen that the optimal type in each group of the 8 groups covers three categories: Drama, Mystery, and Crime. This indicates that the grouping strategy successfully broke the monopoly of a single type and achieved the matching of users' preferences with the optimal type.

3.2 Cumulative regret comparison analysis

The trend of the curve, as shown in Fig.1, indicates that the cumulative regret of the clustered UCB-C and the ordinary UCB varies significantly with the number of iterations T:

At the low-round stage (T≤2000), the regret gap between the clustering UCB-C and the ordinary UCB algorithms is relatively small. During this stage, the algorithm mainly focuses on exploration and has not accumulated sufficient data. The advantage of the reward estimation in each group has not been fully demonstrated, which is consistent with the conclusion in the structured MAB theory that exploration in the early stage relies on prior information [3].

At the Mid-to-high round stage (T > 2000), the rate of regret growth for the clustering UCB-C significantly slows down. This result indicates that the clustering strategy enhances the accuracy of reward distribution estimation by gathering the rating data of homogeneous users. The arm selection during the exploitation phase is closer to the optimal choice within the group, significantly reducing the cost of ineffective exploration [5][10].

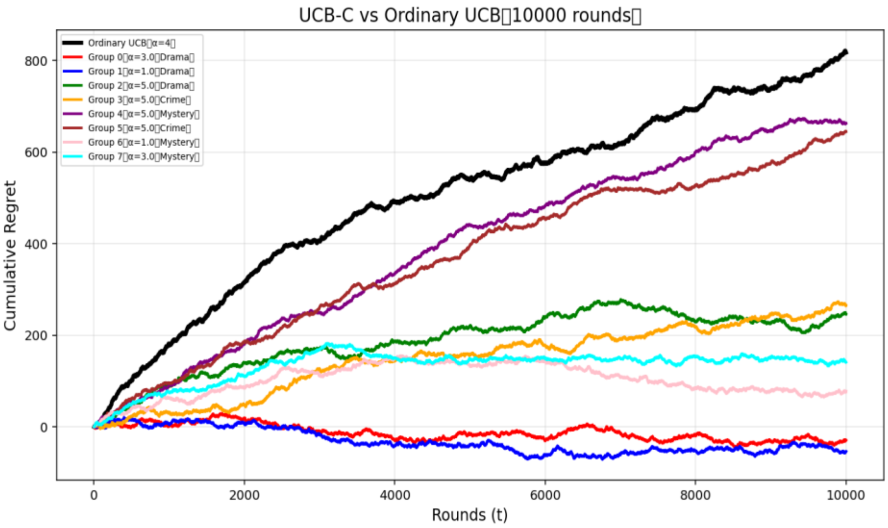

Fig.1. Regret value curves of clustered UCB-C and Ordinary UCB(Picture credit: Original)

4 Discussion and Suggestions

This study used a two-step fixed clustering strategy of "preferences and attributes". The fixed clustering failed to precisely match the heterogeneity of preferences. This limitation is consistent with the inherent defect of traditional K-Means clustering, which "requires the pre-determination of the number of clusters" [6], lacking the flexibility of adaptive clustering algorithms. One possible approach is to replace the fixed K-value KMeans algorithm with a density-based adaptive clustering algorithm. This can be achieved by automatically determining the optimal number of clusters through global and local density estimation [6].

5 Conclusion

This paper aims to achieve low regret recommendations in heterogeneous recommendation systems. Based on the MovieLens 1M dataset, a two-step clustering strategy of "preference clustering and user clustering" and a personalized α allocation mechanism are proposed. The clustering UCB-C algorithm is designed to study the user preference matching and exploration-exploitation balance problem within the framework of structured multi-armed bandit. In this experiment, this paper obtained 8 user groups that covered the three optimal types of "Drama", "Mystery", and "Crime", which satisfied the core assumption of "homogeneous preferences within groups and heterogeneous preferences between groups", thereby verifying the rationality of the clustering strategy. The selection of competing arms effectively controlled the scale of competing arms in each group, reduced ineffective exploration, and provided support for the low regret performance of the clustered UCB-C. Based on the variance of group preferences and average age, three levels of exploration intensity (high, medium, and low) were designed, which can adapt to the exploration needs of different user groups. Finally, the cumulative regret of the clustered UCB-C and the ordinary UCB decreased significantly with the number of iterations.In the future, density-adaptive clustering optimization can be introduced to adjust the granularity of clustering, and a multi-objective optimization framework can be expanded to consider the diversity of recommendations. This research provides a reusable framework for enhancing the practicality of recommendation systems in heterogeneous scenarios, and it can be further applied to e-commerce and advertising fields.

References

[1] Zeng, J.: A comparative study of multi-armed bandit algorithms for dynamic filter and effect recommendations on Tiktok. ITM Web of Conferences. 78, p. 01037 (2025)

[2] He, Q.: Enhancing movie recommendations through comparative analysis of UCB algorithm variants. Applied and Computational Engineering. 68, pp. 45–53 (2024)

[3] Gupta, S., Chaudhari, S., Mukherjee, S., Joshi, G. & Yağan, O.: A unified approach to translate classical bandit algorithms to the structured bandit setting. IEEE Journal on Selected Areas in Information Theory. 1(3), pp. 840–853 (2020)

[4] Gupta, S., Chaudhari, S., Joshi, G. & Yağan, O.: Multi-armed bandits with correlated arms. IEEE Transactions on Information Theory. pp. 1–1 (2021)

[5] Gore, A. & Chaporkar, P.: Clus-UCB: A near-optimal algorithm for clustered bandits. arXiv preprint arXiv:2508.02909 (2025a)

[6] Ban, Y. & He, J.: Local clustering in contextual multi-armed bandits. Proceedings of the Web Conference 2021. pp. 2335–2346 (2021)

[7] Harper, F.M. & Konstan, J.A.: The movielens datasets. ACM Transactions on Interactive Intelligent Systems. 5(4), pp. 1–19 (2015)

[8] Chen, Y.: A comparative study of UCB and Thompson sampling with structured rewards: parameter sensitivity and robustness. ITM Web of Conferences. 80, p. 02004 (2025b)

[9] Qin, M.: Comparative analysis of typical UCB algorithm performance. Applied and Computational Engineering. 46, pp. 131–135 (2024)

[10] Varude, C., Chaudhary, J., Kaushik, S. & Chaporkar, P.: Threshold-based optimal arm selection in monotonic bandits: regret lower bounds and algorithms. arXiv preprint arXiv:2509.02119 (2025)