Contextual Meta-Bandit for Recommender Systems Selection

Marlesson R. O. Santana [1] Luckeciano C. Melo [2] Fernando H. F. Camargo [1] Bruno Brandão [1]
Anderson Soares [1] Renan M. Oliveira [2] Sandor Caetano [2]

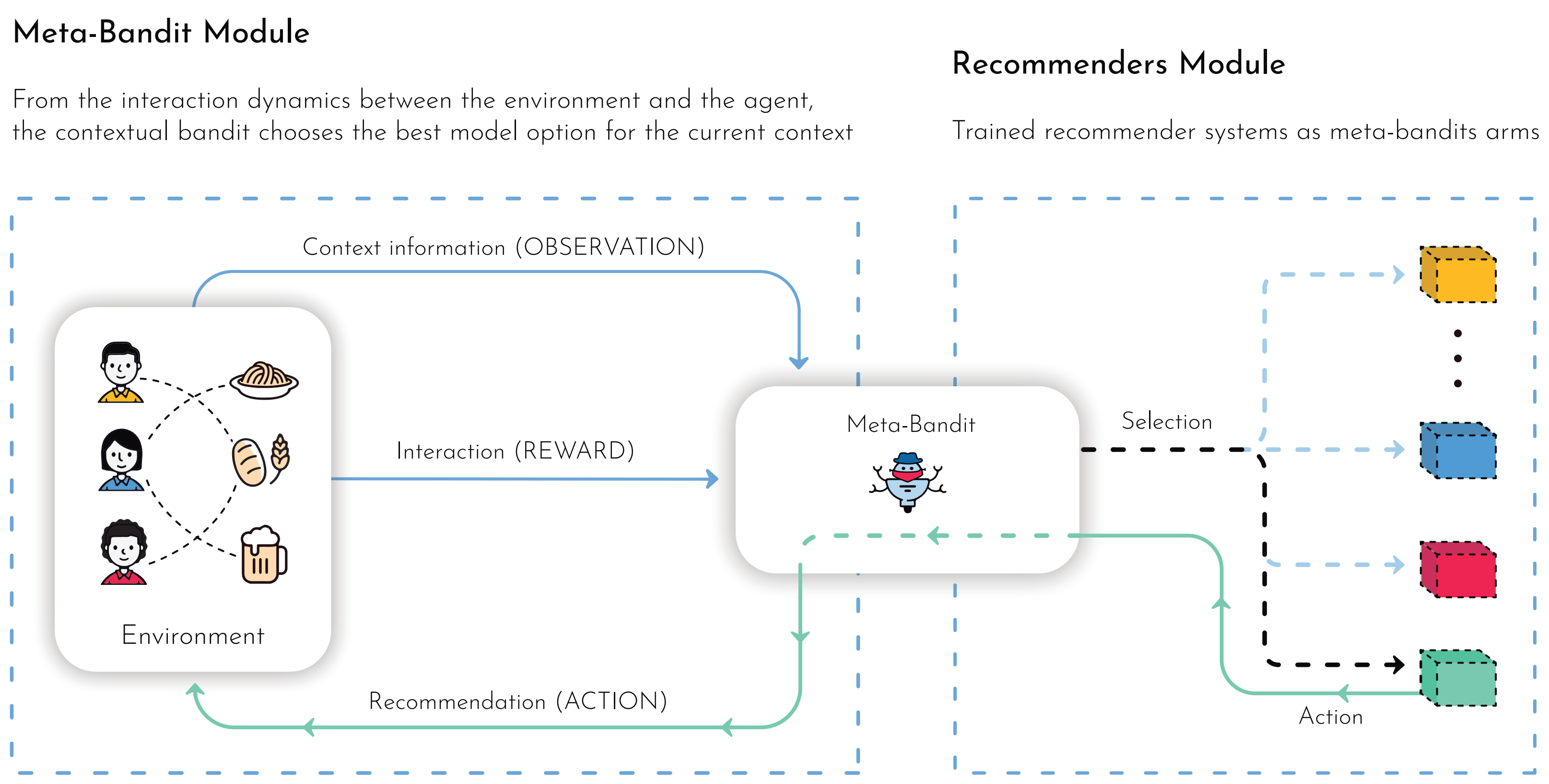
[1]Deep Learning Brazil -- Federal University of Goiás [2]Ifood Research

Introduction

Recommendation systems operate in a highly stochastic and non-stationary environment. As the amount of user-specific information varies, the users' interests themselves also change [1]. This combination creates a dynamic setting where a single solution will rarely be optimal unless it can keep up with these transformations [2]. One system may perform better than others depending on the situation at hand, thus making the choice of which system to deploy, even more difficult. We address these problems by using the Hierarchical Reinforcement Learning framework. Our proposed meta-bandit acts as a policy over options, where each option maps to a pre-trained, independent recommender system. This meta-bandit learns online and selects a recommender accordingly to the context, adjusting to the situation.

Methodology

Figure 1 presents the proposed approach, composed of two modules. Some pre-trained models compose the Recommender Module. These models can be of any type and work as black boxes. They receive an observation as input and recommend items as output. The Meta-Bandit Module works as a proxy, handling the incoming requests. According to the context, it selects which model is more appropriate to provide the recommendations.



Meta-Bandit Module

From the interaction dynamics between the environment and the agent, the contextual bandit chooses the best model option for the current context

Recommenders Module

Trained recommender systems as meta-bandits arms

Figure 1:For each interaction, the environment provides an observation (contextual information). The Meta-Bandit uses it to select one of the recommenders and let the selected one decide the action (recommender item). The environment receives this action and gives a reward to the Meta-Bandit.

In terms of the Semi-MDP, we use $\pi_\Omega : \mathcal{S} \times \Omega \to [0,1]$ to denote a meta-policy, i.e. a policy π over options $\omega \in \Omega$. During the optimization, we seek for a meta-policy that maximizes the reward throughout the episode:

$$\max_{\pi_\Omega} \mathbb{E}_{\pi_\Omega}\Big[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\Big], a_t \sim \overline{\pi}_\omega(\cdot \mid s_t), \overline{\pi}_\omega \sim \pi_\Omega(\cdot \mid s_t), \qquad (1)$$

where γ is the MDP discount factor and T is the length of a training episode. Additionally, we denote $\overline{\pi}_\omega$ to indicate that our work considers fixed intra-policies. Hence, the gradients are not computed through them, which is essential to consider each recommender system as a black-box decision-making procedure.

Experiments

We applied Upper Confidence Bound (UCB), ϵ-Greedy, and Softmax Explorer as exploration strategies for the meta-bandit. We also implemented a classic ensemble for comparison. In this method, we average the recommended items in order to create a combined recommendation from the same algorithms used as arms in the meta-bandit.

We performed the simulation of each proposed baseline model in the test data to observe cumulative mean reward throughout the simulation. For statistical significance, we represent results by the mean and standard deviation across ten executions of the same experiment. In Table 1, we present the results for each experiment.

Table 1:Mean and Cumulative Reward on the test data.

	Mean Reward		Cum. Reward	
	mean	std	mean	std
Random	0.027	-	346	-
Most Popular	0.091	-	1049	-
CDAE	0.094	-	1079	-
Ensemble (MP, MF, CDAE, CVAE)	0.100	-	1146	-
Matrix Factorization	0.105	-	1211	-
CVAE	0.110	-	1272	-
Without Context				
Meta-Bandit UCB[1] (c = 10.0)	0.127	-	1452	-
Meta-Bandit ϵ-greedy[1] (ϵ = 0.2)	0.108	0.003	1243	32
With Context				
Meta-Bandit ϵ-greedy (ϵ = 0.1)	0.155	0.004	1780	45
Meta-Bandit Softmax (c = 500.0)	**0.165**	0.002	**1891**	23

[1]Meta-Bandit without contextual information.

Figure 2 shows that the Most Popular model alternates performance over time. The peaks shown in the figure are correlated with the interactions performed at lunch time, around 11 am. While the declines are the interactions around dinner at 8 pm each day. The remainder methods, including the ensemble, maintain a less variant performance. Meanwhile, the contextual meta-bandit achieves higher rewards over time, outperforming all base-models.

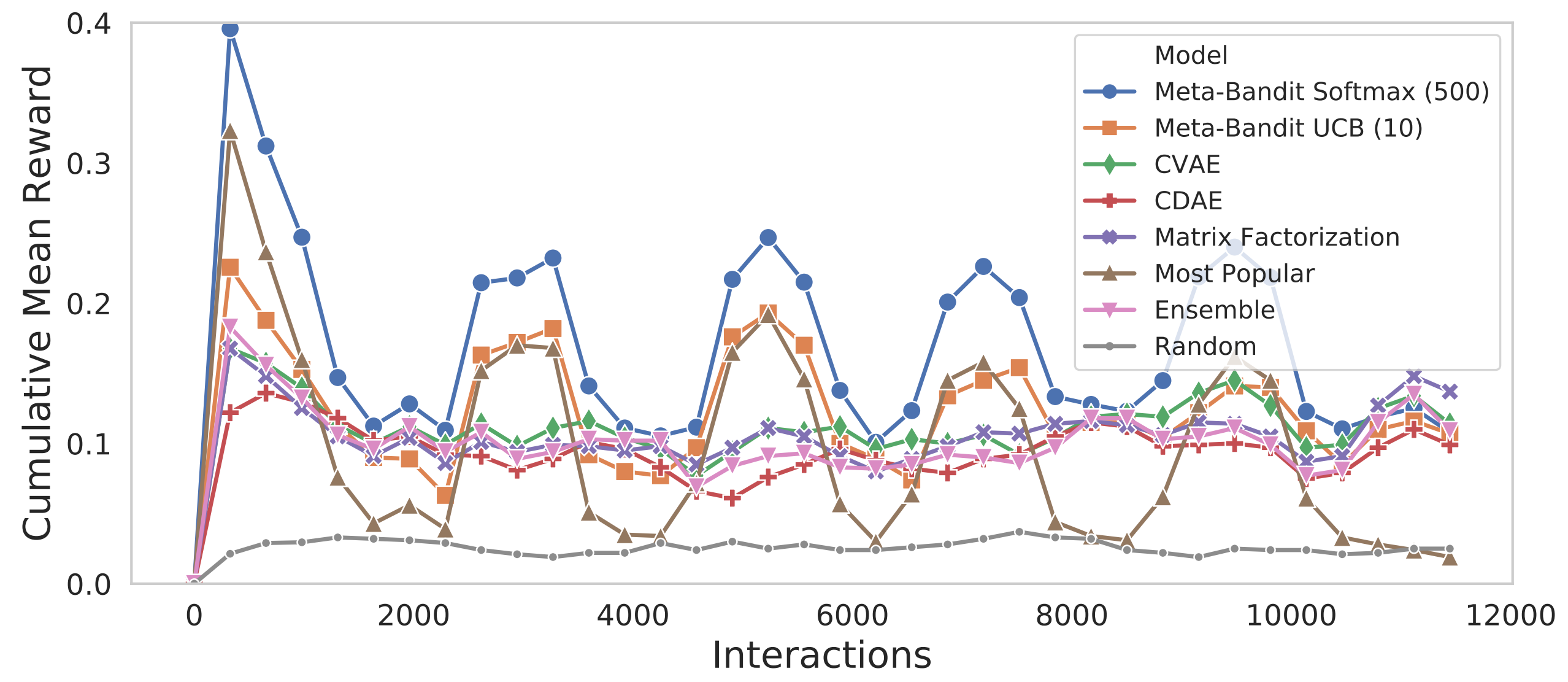
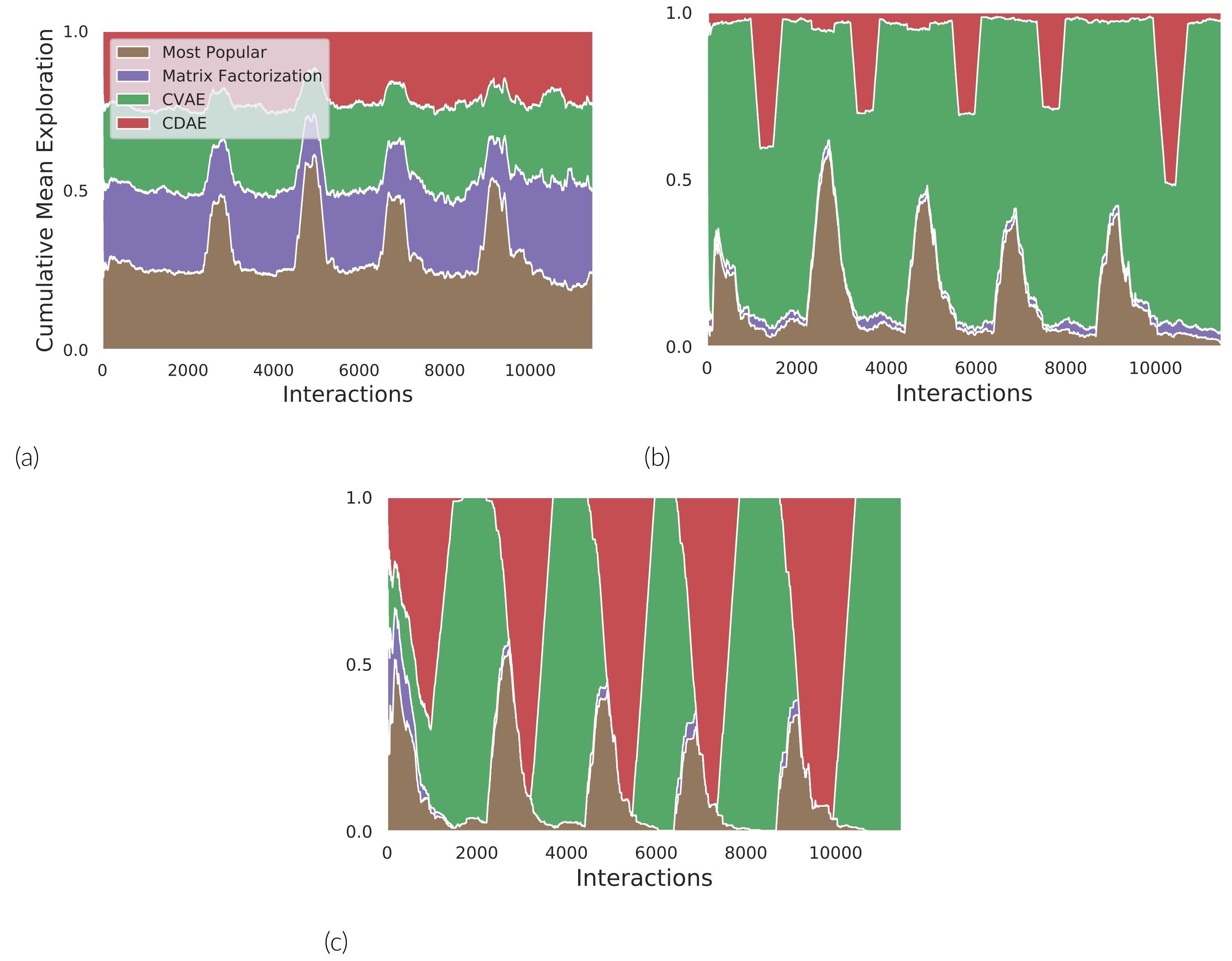


Figure 2:Mean reward from the last 500 interactions

In order to understand the behavior of our meta-bandit in detail, we present the traffic exploration ration of arms. Figure 1 shows how the preferred arm changes over time in cyclic manner. We can also see that the meta-bandit has its decision boundary well defined with the contextual softmax explorer in Figure 1-c.



(a)

(b)

(c)

Figure 3:Interaction over time (x axis) vs. traffic ratio exploration of arms from the last 500 interactions (y axis). **a)** Meta-Bandit with UCB exploration (without context). **b)** Meta-Bandit ϵ-greedy. **c)** Meta-Bandit Softmax Exploration.

Conclusions

In this work, we proposed a meta-bandit approach for Recommender Systems selection, which managed to switch among base-models dynamically, without always favoring one above all others. This behavior leads us to conclude that online learning is beneficial to developing an adaptive recommender. Moreover, we ablated over the usage of contextual information to highlight its importance in reducing the uncertainty over the selection. By having access to sufficient information, our model selector associated its arms to the context changes. All meta-bandits increased the performance above any of its arms alone.

References

[1] Naman Shukla, Arinbjörn Kolbeinsson, Lavanya Marla, and Kartik Yellepeddi.
Adaptive Model Selection Framework: An Application to Airline Pricing.
arXiv:1905.08874 [cs, stat], May 2019.
arXiv: 1905.08874.

[2] Qingyun Wu, Huazheng Wang, Yanen Li, and Hongning Wang.
Dynamic Ensemble of Contextual Bandits to Satisfy Users' Changing Interests.
In *The World Wide Web Conference on - WWW '19*, pages 2080--2090, San Francisco, CA, USA, 2019. ACM Press.

14th ACM Conference on Recommender Systems Online, Worldwide, 22nd-26th September 2020