2ZM-07

A study for the exploration–exploitation strategy of human based on restless two-armed bandit task

Jiaxing Tian[1], Chie Hieida[1], Junichiro Yoshimoto[1,2], Kenta Kimura[3], Hideki Ohira[4], and Kazushi IKEDA [1]

[1]Nara Institute of Science and Technology
[2]Fujita Health University
[3]National Institute of Advanced Industrial Science and Technology
[4]Nagoya University

1 Introduction

Studying human decision-making mechanisms and modeling them can understand and predict decision-making behaviors. A good decision making model can be applied to many studies, such as treating mental illness, game AI, market decision making. As humans live in unstationary environments, in the study of human decision making behavior, exploration–exploitation trade-off is the core issue. Some studies use rewards or punishments as stimuli to study the brain areas related to the exploitation, and they model these function areas with reinforcement learning and get a good fitting results [1, 2]. However, these studies lack the discussion of the exploration–exploitation trade-off mechanism in the models. Some studies have proposed exploration and exploitation are two separate processes of the brain [3]. In order to clarify the exploration–exploitation trade-off mechanism in the human decision-making behavior model, in this study, we divide the coding process into two parts: value function through reward and strategies balance the exploitation–exploration process based on the value function. We investigate decision-making in a restless two-armed bandit task and use multiple methods for each part to fit the dataset. We use Akaike information criterion (AIC) and Bayesian information criterion (BIC) to evaluate the best method and find that the exploration–exploitation trade-off parameters can better classify the human choice patterns.

2 Method

2.1 Materials

We investigated decision making in a restless two-armed bandit task. In this task, 71 healthy adults participated in the study. Each participant made 240 trials of selection. In each trial of the experiment, there are two choices on the screen. According to the choice, there is a probability of getting a reward or a penalty. After every 40 trials of the experiment, the probability of getting a reward or penalty was reversed. Participants asked for the most rewards, and they did not know the probability changing pattern of getting rewards.

Table 1: Compared models

Value function	Selection strategy
Q-learning [4]	ϵ-greedy [5]
Beta distribution [6]	Boltzmann [7]
Bayesian inference [8]	Upper Confidence
Decay [9]	Bound [10]

2.2 Compared models

We divide the coding process into two parts. One part calculates the value function through reward, and the other part of the process balances the exploitation and exploration process based on the value function. We use four models that perform well in the 2-armed bandit problem for the value function part and three methods for the selection strategy part (Table 1). We use these 12 (4 value function x 3 selection strategy) models to fit the dataset. Based on the results of human selection, the model parameters are trained through the maximum likelihood function. We use AIC or BIC to evaluate the best coding method.

3 Result and Discussion

Table 2 contains the fit result for the twelve models. We see the best-fitting model is decay learning with a Upper Confidence Bound (UCB) selection strategy. The fitting results of the Beta distribution are much worse than other value function models. Because in the beta distribution model, we assume the probability of reward as the value function. People usually do not make choices by calculating probabilities. On the other hand, we find that no selection strategy model will always be better than other models. Different value functions need to adopt an appropriate selection strategy to get the best fitting results.

From the result of the experiment, among the 71 participants, each had a different selection strategy

Table 2: Fitting result for the twelve models (lower scores are better)

Value function	Strategies	LLH	AIC	BIC
Q-learning [4]	ϵ-greedy [5]	-10546.56	21519.13	21105.96
	Boltzmann [7]	-10472.66	21371.32	20958.15
	UCB [10]	-10550.14	21526.28	21113.11
Beta distribution ([6])	ϵ-greedy	-11357.37	23140.74	22727.57
	Boltzmann	-11661.94	23749.88	23336.71
	UCB	-11494.28	23414.56	23001.39
Bayesian inference [8]	ϵ-greedy	-10570.66	21709.32	21158.43
	Boltzmann	-10720.74	22009.49	21458.60
	UCB	-10551.39	21670.78	21119.89
Decay [9]	ϵ-greedy	-10521.70	21469.41	21056.24
	Boltzmann	-10514.17	21454.34	21041.17
	UCB	**-10408.38**	**21242.76**	**20829.59**

based on the value function. Ten people are inclined to exploration (participants tried other options many times after being awarded consecutively), and nine people are inclined to exploitation (participants were unwilling to change their choice after receiving rewards). When the two-class classification was performed using the exploration and exploitation trade-off parameters of decay learning with a UCB, the correct answer rate was 100 %.

4 Conclusion

This study found that the model with the best fitting result (decay learning with a UCB) can distinguish the two groups of participants through the exploration–exploitation trade-off parameter in the selection strategy. Therefore, it is feasible to use the model structure with the exploration–exploitation trade-off method proposed in this study to model the decision making behavior.

If we can model the exploration-exploitation trade-off parameter by analyzing fMRI data, then we can use this to analyze human decision making behavior.

Acknowledgement

This work was supported by JSPS KAKENHI Grant-in-Aid for Early-Career Scientists (20K19907).

References

[1] Makoto Ito and Kenji Doya. “Validation of decision-making models and analysis of decision variables in the rat basal ganglia”. In: *Journal of Neuroscience* 29.31 (2009), pp. 9861–9874.

[2] Atsuo Yoshino et al. “Importance of the habenula for avoidance learning including contextual cues in the human brain: A preliminary fMRI study”. In: *Frontiers in Human Neuroscience* 14 (2020), p. 165.

[3] Daniella Laureiro-Martinez et al. “Understanding the exploration–exploitation dilemma: An fMRI study of attention control and decision-making performance”. In: *Strategic management journal* 36.3 (2015), pp. 319–338.

[4] Christopher JCH Watkins and Peter Dayan. “Q-learning”. In: *Machine learning* 8.3-4 (1992), pp. 279–292.

[5] Richard S Sutton, Andrew G Barto, et al. *Introduction to reinforcement learning*. Vol. 135. MIT press Cambridge, 1998.

[6] Ole-Christoffer Granmo. “Solving two-armed Bernoulli bandit problems using a Bayesian learning automaton”. In: *International Journal of Intelligent Computing and Cybernetics* (2010).

[7] Kavosh Asadi and Michael L Littman. “An alternative softmax operator for reinforcement learning”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 243–252.

[8] Rudolph Emil Kalman. “A new approach to linear filtering and prediction problems”. In: (1960).

[9] Ido Erev and Alvin E Roth. “Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria”. In: *American economic review* (1998), pp. 848–881.

[10] Wassim Jouini et al. “Upper confidence bound based decision making strategies and dynamic spectrum access”. In: *2010 IEEE International Conference on Communications*. IEEE. 2010, pp. 1–5.