Adaptive Exploration in Stochastic Multi-armed Bandit Problem

Qian Zhou¹, Xiaofang Zhang¹,², Peng Zhang¹, Quan Liu¹,³,⁴

1. School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu, 215006
2. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023
3. Collaborative Innovation Center of Novel Software Technology and Industrialization, Nanjing, 210000
4. Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, 130012.

* Corresponding author email: xfzhang@suda.edu.cn

Abstract:

The Multi-Armed Bandit (MAB) problem is a typical problem of the exploration and exploitation dilemma in reinforcement learning. As a classical MAB problem, the Stochastic Multi-Armed Bandit (SMAB) problem is the base of many new MAB problems. To solve the problems of insufficient use of information in existing SMAB methods, this paper presents an adaptive algorithm to balance exploration and exploitation based on the Chosen Number of Arm with Minimal Value, namely CNAMV in short. The upper bound of CNAMV’s regret was theoretically proved, and our experimental results showed that CNAMV could yield greater reward and smaller regret with high efficiency than commonly used methods. Therefore, CNAMV is a cost-effective SMAB method.

Conclusions

In this paper, we proved the upper bound of CNAMV’s regret theoretically, and discussed the influence of the key parameter w in CNAMV algorithm. Through three experiments, we provided the reference range of parameter w and compared CNAMV with classical algorithms and their variants in the random data set and the content distribution network dataset respectively. The experimental results showed that the CNAMV algorithm could yield greater reward and smaller regret than ε-greedy, ε-decreasing, SoftMax, decreasing SoftMax and UCB1. As a result, CNAMV algorithm is a cost-effective stochastic multi-armed bandit algorithm.

In future work, we intend to extend the CNAMV algorithm to more complex settings, such as budgeted multi-armed bandits and qualitative multi-armed bandits, and put forward practical application of new multi-armed bandit problem.

References


Acknowledgements

Here and now, I would like to extend my sincere thanks to all those who have helped me make this thesis possible and better. Firstly, I am deeply grateful to my honorable supervisor, Xiaofang Zhang, who has checked through my thesis with patience and given me instructive suggestions, and she also played an important role in indicating a bright road in my future writing. Then thanks to the teachers and professors who have taught me. Finally, I am very grateful to my lovely friends and classmates who have offered me quiet situation to compose my thesis and discussed with me about my thesis.