MOL2NET, 2016, Vol. 2, J; http://sciforum.net/conference/MOL2NET-02/SUIWML-01

SciForum
MOL2NET

Adaptive Exploration in Stochastic Multi-armed Bandit Problem

Qian Zhou[1], Xiaofang Zhang[1,2], Peng Zhang[1], Quan Liu[1,3,4]

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

[1] Sutton R, Barto G. Reinforcement learning: an introduction [M]. Cambridge, MA: the MIT Press. 1998.
[2] Robbins H. Some aspects of the sequential design of experiments [J]. American Mathematical Society. 1952, 55: 527-535.
[3] Devanur N R, Kakade S M. The price of truthfulness for pay-per-click auctions[C]// ACM Conference on Electronic Commerce. 2009:99-106.

[4] Jain S, Gujar S, Zoeter O. A quality assuring multi-armed bandit crowdsourcing mechanism with incentive compatible learning [J]. Carbon, 2014:1609-1610.
[5] Jain S, Narayanaswamy B and Narahari Y. A multi-armed bandit incentive mechanism for crowdsourcing demand response in smart grids[C]//Proceedings of the twenty-eighth AAAI conference on artificial intelligence and the twenty-sixth innovative applications of artificial intelligence conference, 2014: 721-727.
[6] Bubeck S, Cesabianchi N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems [J]. Foundations & Trends® in Machine Learning, 2012, 5(1):1-122.
[7] Watkins C J C H. Learning from delayed rewards [J]. Robotics & Autonomous Systems, 1989, 15(4):233-235.
[8] Luce R D. Individual choice behavior [J]. American Economic Review, 1959, 67(1):1-15.
[9] Lahaie S. An analysis of alternative slot auction designs for sponsored search[C]//Proceedings of the 7th ACM Conference on Electronic Commerce. ACM, 2006:218-227.
[10] Krishnamurthy B, Wills C, Zhang Y. On the use and performance of content distribution networks [C]//Proceedings of the First ACM Sigcomm Internet Measurement Workshop, 2001:169-182.
[11] Thathachar M A L, Sastry P S. A new approach to the design of reinforcement schemes for learning automata [J]. IEEE Transactions on Systems Man & Cybernetics, 1985, SMC-15(1):168-175.
[12] Even-Dar E, Mannor S, Mansour Y. PAC bounds for multi-armed bandit and markov decision processes[C]//In Fifteenth Annual Conference on Computational Learning Theory (COLT), 2002:255-270.
[13] Cesa-Bianchi N, Fischer P. Finite-time regret bounds for the multi-armed bandit problem[C]//ICML. 1998:100-108.
[14] Strens M. Learning Cooperation and feedback in pattern recognition [D]. Physics Department, King’s College London, 1999.
[15] Vermorel J, Mohri M. Multi-armed bandit algorithms and empirical evaluation[C]// European Conference on Machine Learning. 2005:437-448.
[16] Kirkpatrick B S, Gelatt C, Vecchi D. Optimization by simulated annealing [J]. Science, 1983, 220(4598):671-680.
[17] Auer P, Cesa-Bianchi N, Fischer P. Finite-time analysis of the multiarmed bandit problem[J]. Machine Learning, 2002, 47(2-3):235-256.
[18] Lai T L, Robbins H. Asymptotically efficient adaptive allocation rules [J]. Advances in Applied Mathematics, 1985, 6(1):4-22.
[19] Xia Y, Li H, Qin T. Thompson sampling for budgeted multi-armed bandits[C]// International Conference on Artificial Intelligence. AAAI Press, 2015:3960-3966.
[20] Szörényi B, Busa-Fekete R, Weng P. Qualitative multi-armed bandits: a quantile-based approach[C]//International Conference on Machine Learning. 2015:1660-1668.

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.