SciForum
MOL2NET

Single Trajectory Learning: Exploration VS. Exploitation

Qiming Fu [1,3,4], Quan Liu [2,5] , Heng Luo [1,3,4] and Jianping Chen [1,3,4,]*

[1] Institute of Electronics and Information Engineering, Suzhou University of Science and Technology, Suzhou, Jiangsu; E-Mail: fqm_1@mail.usts.edu.cn; hengluo@mail.usts.edu.cn; alanjpchen@yahoo.com

[2] School of Computer Science and Technology, Soochow University, Suzhou, Jiangsu; E-Mail: quanliu@suda.edu.cn

[3] Jiangsu Province Key Laboratory of Intelligent Building Energy Efficiency, Suzhou University of Science and Technology, Suzhou, Jiangsu

[4] Suzhou Key Laboratory of Mobile Network Technology and Application, Suzhou University of Science and Technology, Suzhou, Jiangsu

[5] Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun

Abstract: *In reinforcement learning, the exploration/exploitation dilemma is a very crucial issue, which can be described as searching between the exploration of the environment to find more profitable actions, and the exploitation of the best empirical actions for the current state. We focus on the single trajectory reinforcement learning problem where an agent is interacting with a partially unknown MDP over single trajectories, and try to deal with the exploration/exploitation in this setting. Given the reward function, we try to find a good E/E strategy to address the MDPs under some MDP distribution. This is achieved by selecting the best strategy in mean over a potential MDP distribution from a large set of candidate strategies, which is done by exploiting single trajectories drawn from plenty of MDPs. In this paper, we mainly make the following contributions: 1) we discuss the strategy-selector algorithm based on formula set and polynomial function.2) we provide the theoretical and experimental regret analysis of the learned strategy under an given MDP distribution. 3) we compare these methods with the ``state-of-the-art" Bayesian RL method experimentally.*

Keywords: *single trajectory, MDP distribution, E/E delimma, Bayesian reinforcement learning*