通信学报
Journal on Communications
主管单位：中国科学技术协会
主办单位：中国通信学会

于俊1,刘全1,2,傅启明1,孙洪坤1,陈桂兴1.基于优先级扫描Dyna结构的贝叶斯Q学习方法[J].通信学报,2013,(11):129~139

基于优先级扫描Dyna结构的贝叶斯Q学习方法

Bayesian Q learning method with Dyna architecture and prioritized sweeping

投稿时间： 2013-05-18

DOI：10.3969/j.issn.1000-436x.2013.11.015

中文关键词：强化学习 马尔科夫决策过程 优先级扫描 Dyna结构 贝叶斯Q学习

英文关键词:reinforcement learning Markov decision process prioritized sweeping Dyna architecture Bayesian Q learning

基金项目:国家自然科学基金资助项目(61070223, 61103045, 61070122, 61272005)；江苏省自然科学基金资助项目(BK2012616)；江苏省高校自然科学研究基金资助项目(09KJA520002, 09KJB520012)；吉林大学符号计算与知识工程教育部重点实验室基金资助项目(93K172012K04)

作者	单位
于俊1，刘全1,2，傅启明1，孙洪坤1，陈桂兴1	1. 苏州大学 计算机科学与技术学院，江苏 苏州 215006；2. 吉林大学 符号计算与知识工程教育部重点实验室，吉林 长春 130012

中文摘要:

贝叶斯Q学习方法使用概率分布来描述Q值的不确定性，并结合Q值分布来选择动作，以达到探索与利用的平衡。然而贝叶斯Q学习存在着收敛速度慢且收敛精度低的问题。针对上述问题，提出一种基于优先级扫描Dyna结构的贝叶斯Q学习方法—Dyna-PS-BayesQL。该方法主要分为2部分：在学习部分，对环境的状态迁移函数及奖赏函数建模，并使用贝叶斯Q学习更新动作值函数的参数；在规划部分，基于建立的模型，使用优先级扫描方法和动态规划方法对动作值函数进行规划更新，以提高对历史经验信息的利用，从而提升方法收敛速度及收敛精度。将Dyna-PS-BayesQL应用于链问题和迷宫导航问题，实验结果表明，该方法能较好地平衡探索与利用，且具有较优的收敛速度及收敛精度。

英文摘要:

In order to balance this trade-off, a probability distribution was used in Bayesian Q learning method to describe the uncertainty of the Q value and choose actions with this distribution. But the slow convergence is a big problem for Bayesian Q-Learning. In allusion to the above problems, a novel Bayesian Q learning algorithm with Dyna architecture and prioritized sweeping, called Dyna-PS-BayesQL was proposed. The algorithm mainly includes two parts: in the learning part, it models the transition function and reward function according to collected samples, and update Q value function by Bayesian Q-learning, in the programming part, it updates the Q value function by using prioritized sweeping and dynamic programming methods based on the constructed model, which can improve the efficiency of using the historical information. Applying the Dyna-PS-BayesQL to the chain problem and maze navigation problem, the results show that the proposed algorithm can get a good performance of balancing the exploration and exploitation in the learning process, and get a better convergence performance.

查看全文 查看/发表评论 下载PDF阅读器

关闭