**强化学习进展调研**

沈正圆 2111812118 计算机技术1802

该调研报告主要利用知网，IEEE/IET和ACM几个学术网站的搜索工具查阅文献后对强化学习的进展做一个深入了解。

### 3篇综述性论文：

中国知网：

[1]高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004, 30(1).

IEEE/IET：

[2]Grondman I , Bucsoniu L , Lopes G A D , et al. A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients[J]. IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), 2012, 42(6):1291-1307.

ACM:

[3] Dusparic I , Cahill V . Autonomic multi-policy optimization in pervasive systems[J]. ACM Transactions on Autonomous and Adaptive Systems, 2012, 7(1):1-25.

### 最具代表性的5篇论文摘要：

中国知网：

[1]高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004, 30(1).

**摘要**：强化学习通过试错与环境交互获得策略的改进,其自学习和在线学习的特点使其成为机器学习研究的一个重要分支.该文首先介绍强化学习的原理和结构;其次构造一个二维分类图,分别在马尔可夫环境和非马尔可夫环境下讨论最优搜索型和经验强化型两类算法;然后结合近年来的研究综述了强化学习技术的核心问题,包括部分感知、函数估计、多agent强化学习,以及偏差技术;最后还简要介绍强化学习的应用情况和未来的发展方向.

[2] 陈学松, 杨宜民. 强化学习研究综述[J]. 计算机应用研究, 2010, 27(8):2834-2838.

**摘要**：在未知环境中,关于agent的学习行为是一个既充满挑战又有趣的问题,强化学习通过试探与环境交互获得策略的改进,其学习和在线学习的特点使其成为机器学习研究的一个重要分支。介绍了强化学习在理论、算法和应用研究三个方面最新的研究成果,首先介绍了强化学习的环境模型和其基本要素;其次介绍了强化学习算法的收敛性和泛化有关的理论研究问题;然后结合最近几年的研究成果,综述了折扣型回报指标和平均回报指标强化学习算法;最后列举了强化学习在非线性控制、机器人控制、人工智能问题求解、多agent系统问题等若干领域的成功应用和未来的发展方向

IEEE/IET：

[3] Sutton R S , Barto A G . Reinforcement Learning: An Introduction[J]. IEEE Transactions on Neural Networks, 1998, 9(5):1054-1054.

**Abstract**: ABSTRACT Of several responses made to the same situation, those which are accompanied or closely followed by satisfaction to the animal will, other things being equal, be more firmly connected with the situation, so that, when it recurs, they will be more likely to recur; those which are accompanied or closely followed by discomfort to the animal will, other things being equal, have their connections with that situation weakened, so that, when it recurs, they will be less likely to occur. The greater the satisfaction or discomfort, the greater the strengthening or weakening of the bond. (Thorndike, 1911) The idea of learning to make appropriate responses based on reinforcing events has its roots in early psychological theories such as Thorndike's "law of effect" (quoted above). Although several important contributions were made in the 1950s, 1960s and 1970s by illustrious luminaries such as Bellman, Minsky, Klopf and others (Farley and Clark, 1954; Bellman, 1957; Minsky, 1961; Samuel, 1963; Michie and Chambers, 1968; Grossberg, 1975; Klopf, 1982), the last two decades have wit- nessed perhaps the strongest advances in the mathematical foundations of reinforcement learning, in addition to several impressive demonstrations of the performance of reinforcement learning algo- rithms in real world tasks. The introductory book by Sutton and Barto, two of the most influential and recognized leaders in the field, is therefore both timely and welcome. The book is divided into three parts. In the first part, the authors introduce and elaborate on the es- sential characteristics of the reinforcement learning problem, namely, the problem of learning "poli- cies" or mappings from environmental states to actions so as to maximize the amount of "reward".

[4] Berenji H R , Khedkar P . Learning and tuning fuzzy logic controllers through reinforcements[J]. IEEE Transactions on Neural Networks, 1992, 3(5):724.

**Abstract**: A method for learning and tuning a fuzzy logic controller based on reinforcements from a dynamic system is presented. It is shown that: the generalized approximate-reasoning-based intelligent control (GARIC) architecture learns and tunes a fuzzy logic controller even when only weak reinforcement, such as a binary failure signal, is available; introduces a new conjunction operator in computing the rule strengths of fuzzy control rules; introduces a new localized mean of maximum (LMOM) method in combining the conclusions of several firing control rules; and learns to produce real-valued control actions. Learning is achieved by integrating fuzzy inference into a feedforward network, which can then adaptively improve performance by using gradient descent methods. The GARIC architecture is applied to a cart-pole balancing system and demonstrates significant improvements in terms of the speed of learning and robustness to changes in the dynamic system's parameters over previous schemes for cart-pole balancing.

ACM:

[5] Mnih V , Kavukcuoglu K , Silver D , et al. Playing Atari with Deep Reinforcement Learning[J]. Computer Science, 2013.

**Abstract**: We present the first deep learning model to successfully learn control policies directly from high-dimensional sensory input using reinforcement learning. The model is a convolutional neural network, trained with a variant of Q-learning, whose input is raw pixels and whose output is a value function estimating future rewards. We apply our method to seven Atari 2600 games from the Arcade Learning Environment, with no adjustment of the architecture or learning algorithm. We find that it outperforms all previous approaches on six of the games and surpasses a human expert on three of them.

### 该主题研究方向几个较有代表性的国内和国外研究学者及其单位：

1. DeepMind

位于英国伦敦，是由人工智能程序师兼神经科学家戴密斯·哈萨比斯(Demis Hassabis)等人联合创立，是前沿的人工智能企业，其将机器学习和系统神经科学的最先进技术结合起来，建立强大的通用学习算法。最初成果主要应用于模拟、电子商务、游戏开发等商业领域。

目前，Google 旗下的 DeepMind 已经成为 AI 领域的明星，据外媒 2016年6月8日，DeepMind 欲将其算法应用到医疗保健行业，包括计划在 5年内使用机器学习处理英国国家医疗服务体系（以下简称：NHS） 的数据。

1. OpenAI

由诸多硅谷大亨联合建立的人工智能非营利组织。2015年马斯克与其他硅谷科技大亨进行连续对话后，决定共同创建OpenAI，希望能够预防人工智能的灾难性影响，推动人工智能发挥积极作用。特斯拉电动汽车公司与美国太空技术探索公司SpaceX创始人马斯克、Y Combinator总裁阿尔特曼、天使投资人彼得·泰尔（Peter Thiel）以及其他硅谷巨头去年12月份承诺向OpenAI注资10亿美元。

OpenAI2016年6月21日宣布了其主要目标，包括制造“通用”机器人和使用自然语言的聊天机器人。OpenAI研发主管伊利娅·苏特斯科娃(Ilya Sutskever)、OpenAI CTO格雷格·布劳克曼（Greg Brockman）硅谷知名创业加速器Y Combinator总裁萨姆·阿尔特曼（Sam Altman）以及连续创业家伊隆·马斯克（Elon Musk）等人联合发表博文称：“我们正致力于利用物理机器人（现有而非OpenAI开发）完成基本家务。

### 研究进展

1. 什么是强化学习

下面给出强化学习基本组成元素的定义：

智能体：强化学习的本体，作为学习者或者决策者。

环境：强化学习智能体以外的一切，主要由状态集合组成。

状态：一个表示环境的数据，状态集则是环境中所有可能的状态。

动作：智能体可以做出的动作，动作集则是智能体可以做出的所有动作。

奖励：智能体在执行一个动作后，获得的正/负反馈信号，奖励集则是智能体可以获得的所有反馈信息。

策略：强化学习是从环境状态到动作的映射学习，称该映射关系为策略。通俗的理解，即智能体如何选择动作的思考过程称为策略。

目标：智能体自动寻找在连续时间序列里的最优策略，而最优策略通常指最大化长期累积奖励。

因此，强化学习实际上是智能体在与环境进行交互的过程中，学会最佳决策序列。

1. 强化学习研究进展

1959年，人工智能先驱Arthur Samuel正式定义了“机器学习”这概念。也正是这位Samuel，在50年代开发了基于RL的的象棋程序，成为人工智能领域最早的成功案例[63]。为何人工智能先驱们的工作往往集中在RL相关的任务呢？经典巨著《人工智能：一种现代方法》里对RL的评论或许可以回答这一问题：可以认为RL囊括了人工智能的所有要素：一个智能体被置于一个环境中，并且必须学会在其间游刃有余。

2015年，DeepMind的Volodymyr Mnih等研究员在《自然》杂志上发表论文Human-level control through deep reinforcement learning[1]，该论文提出了一个结合深度学习（DL）技术和强化学习（RL）思想的模型Deep Q-Network(DQN)，在Atari游戏平台上展示出超越人类水平的表现。自此以后，结合DL与RL的深度强化学习（Deep Reinforcement Learning, DRL）迅速成为人工智能界的焦点。

### 参考文献

[1]高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004, 30(1).

[2]Grondman I , Bucsoniu L , Lopes G A D , et al. A Survey of Actor-Critic Reinforcement。

[3] Dusparic I , Cahill V . Autonomic multi-policy optimization in pervasive systems[J]. ACM Transactions on Autonomous and Adaptive Systems, 2012, 7(1):1-25.

[4] 陈学松, 杨宜民. 强化学习研究综述[J]. 计算机应用研究, 2010, 27(8):2834-2838.

[5] Mnih V , Kavukcuoglu K , Silver D , et al. Playing Atari with Deep Reinforcement Learning[J]. Computer Science, 2013.

[6] Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540):529-533.

[7] Sutton R S, Precup D, Singh S. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning[J]. Artificial Intelligence, 1999, 112(1-2):181-211.

[8] Holroyd C B, Coles M G. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity.[J]. Psychological Review, 2002, 109(4):679.

[9] Erev I, Roth A E. Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria[J]. American Economic Review, 1998, 88(4):848-881.

[10] Ng A, Russell S. Algorithms for inverse reinforcement learning[C]// International Conference on Machine Learning. 2000.