两种强化学习算法的性能对比

王恩侃[1] 姜 乐[1]
WANG Enkan JIANG Le

摘 要 针对 DQN 和 Dueling-DQN 这两种结合深度学习的强化学习算法，在经验池、学习率和批量采样大小固定的情况下，选取大小分别为 15、30 和 50 的动作空间，在不同复杂程度的游戏环境下进行了算法性能对比。对算法进行初始化后，固定训练步数，对比两种算法的损失值和累计奖励，同时比较达到相同的损失值和累计奖励所需要的训练步数，两种方法相互验证，得出结论。结果表明，Dueling-DQN 算法可以更好地提升算法性能，并且拥有更好的收敛性，其收敛速度相较 DQN 算法提升了 20%。

关键词 强化学习；DQN；Dueling；收敛速度

doi：10.3969/j.issn.1672-9528.2021.07.034

0 引言

强化学习是人工智能的一大领域，其本质就是在决策问题中智能体从环境反馈的知识中学习到一个好的策略的过程[1]。强化学习中包含很多种算法，传统的 Q-learning 算法需要不断更新 Q 值表，只能处理一些简单和低维的问题。而现实中的问题往往都比较复杂，传统算法难以解决。因此 Mnih 等人结合了 Q-learning 算法与神经网络，提出 DQN 算法[2]，解决了传统算法的“维数灾难”问题。Baram 等人根据 DQN 算法基提出了自举 DQN 算法，使用随机值函数来计算，提高了算法的收敛速度[3]。文献 [4] 提出一种基于先验知识的 PK-DQN 改进算法，在 DQN 算法的基础上提高了稳定性。在一些深度学习任务中，特殊状态下，值的大小与动作无关，因此 Wang 等人提出了一种竞争网络算法（Dueling-DQN）[5]。

针对 DQN 和 Dueling-DQN 两种算法，本文使用了同一种模拟环境，更改其行为空间的大小，从而创造出不同复杂程度的环境，通过 cost 值和累计奖励对比两种算法可知 Dueling-DQN 算法的性能获得了极大的提升。

1 DQN 算法

DQN 算法将 Q-learning 算法中维护的 Q 表转化为函数的拟合问题，通过拟合的函数代替 Q 表产生 Q 值，从而预测动作。神经网络具有理论上拟合任意函数的能力，因此 DQN 将神经网络与 Q 学习进行结合，利用神经网络来处理高维复杂的数据。在该算法中引入了两个概念：经验池和 Q-target 网络。经验池中存放着过去学习过程中的一些数据集合，每次更新时，会随机从经验池中选取一些数据集来学习。这样做是为了打断强化学习各个状态之间的相关性，使神经网络训练更有效率。该算法还使用了两个结构相同但参数不同的网络，分别是 Q-target 网络和 Q-eval 网络。Q-eval 使用的是最新的参数，而 Q-target 使用的是以前的参数，每经过一定次数的迭代，便将 Q-eval 的参数复制给 Q-target 网络。Q-target 的使用可以在一段时间内使得目标 Q 值保持不变，降低了当前 Q 值和目标 Q 值之间的联系。其本质上也是打断相关性的一种方法，可以提升算法的稳定性。

DQN 算法流程可用图 1 来表示。

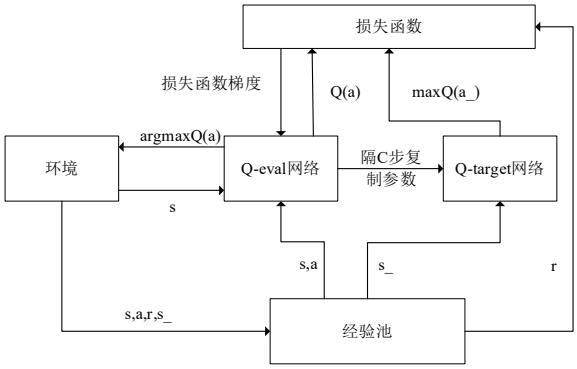

图 1 DQN 算法框图

（1）初始化经验池，两个网络参数及当前状态；

（2）根据当前状态选择动作并获得奖励及下一个状态；将状态，动作，奖励及下一个状态存入经验池；

（3）从经验池中随机采取一些样本学习；

（4）根据公式（1）计算结果；

（5）使用公式（2）作为损失函数来更新网络参数；

（6）每迭代一定次数，将 Q-eval 网络参数复制给 Q-target 网络。

$$Q(s,a)\leftarrow Q(s,a)+\alpha[r+\gamma\max_{a'}Q(s',a')-Q(s,a)] \quad (1)$$

式中：取下一个状态 s' 对应的最大 Q 值 $Q(s',a')$ 乘以衰变 γ，再加上奖励值 r 作为 Q 现实（Q-target 的输出），选取状态

1. 沈阳理工大学信息科学与工程学院 辽宁沈阳 110159

s 所对应的最大 Q 值 $Q(s,a)$ 作为 Q 估计（Q-eval 的输出），以此来不断地更新 Q 值。

$$L(\theta)=E\left[\left(T\arg etQ-Q(s,a,\theta)\right)^2\right] \quad (2)$$

虽然 DQN 算法的应用已经很广泛，但在一些特殊任务中仍然无法令人满意。

2 Dueling-DQN 算法

在一些基于视觉的感知深度强化学习任务中，不同的行为对应不同的值，但在某些状态下，值的大小与行为无关。而 DQN 算法选取最大 Q 值进行决策，网络的更新只依赖于动作值，显然在一些任务中并不适合。改进后的 Dueling-DQN 网络如图 2 所示，该算法将输出分成了两部分，第一部分仅仅与状态有关，而与要采取的行为无关，这部分称为价值函数部分，第二部分与状态和行为都相关，这部分叫作优势函数部分。这种结构的优势在于可以学习到每个状态的值，而不需要考虑在该状态采取什么样的动作。价值函数部分可以用 $V(s,\theta,\beta)$ 来表示，优势函数部分可以用 $A(s,a,\beta,\alpha)$ 来表示。

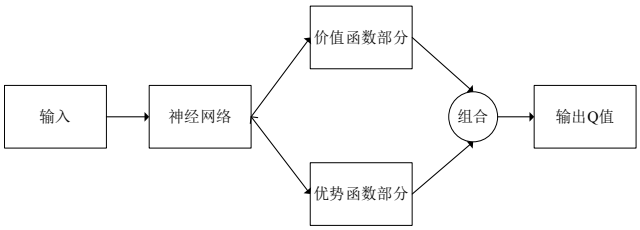

图 2 Dueling-DQN 结构图

动作 Q 值函数公式表示为：

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+A(s,a;\theta,\alpha) \quad (3)$$

式中：θ 是两个部分都存在的网络参数，而 α 是优势函数特有的网络参数，β 是价值函数独有的网络参数。

在实际情况中，公式（3）要将优势函数部分减去某一个状态下所有优势函数部分的均值，该步骤实质上就是对优势函数部分做了中心化处理，可以保证该状态下每个动作相对应的优势函数排序不变。并且缩短了 Q 值范围，提高算法稳定性。

3 对比与分析

Gym 库是 Open AI 推出的强化学习实验环境库，它用 python 语言实现了离散时间智能体 / 环境接口中的环境部分。如图 3 所示，本文使用了 Gym 库中的一个 Atari 游戏 Pendulum-v0（倒立摆）作为仿真的环境部分。

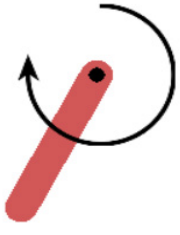

图 3 Pendulum-v0 游戏演示图

在 Pendulum-v0（倒立摆）游戏中，钟摆以一个随机的位置和方向开始向上摆动，直到保持直立。在本次实验中，从经验池批量采样的大小为 32，学习率为 0.001，经验池大小为 3000，动作空间分为 15、30 和 50 三大类，代表了从简单到复杂的三种不同类型的环境。钟摆每成功的直立一次，获得一个 +0 的奖励，其余状态奖励皆为负值。将钟摆摆动的次数作为训练步数，将每次训练获得的奖励求和作为累计奖励。本文将 DQN 算法和 Dueling-DQN 算法进行了对比，通过比较训练步数下的累计奖励和 cost 值来判断改进算法的有效性。若相同的训练步数下累计奖励更大，或 cost 值在较少的训练步数内接近 0，则说明算法的有效性更高。图 4 ～图 6 分别是在不同的行为空间下，两种算法累计奖励和 cost 值的对比。

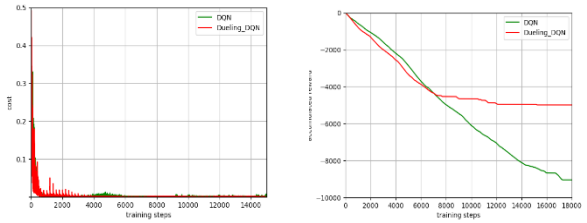

图 4 行为空间为 15 的算法对比图

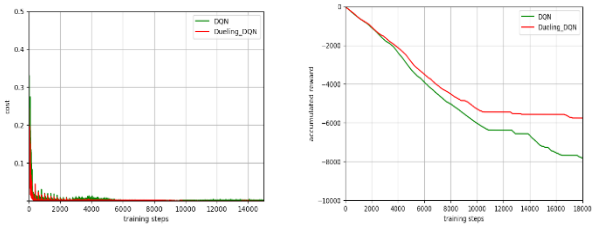

图 5 行为空间为 30 的算法对比图

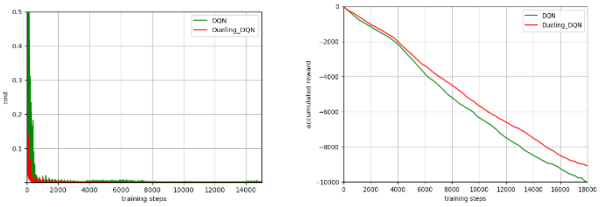

图 6 行为空间为 50 的算法对比图

通过比较图 4 ～图 6 可知，从行为空间 15 ～ 50 的过程中，也就是环境从简单到复杂的一个过程中，在相同的训练步数下，虽然随着环境的复杂，收敛速度也随之变慢，但是 Dueling-DQN 算法总是要比 DQN 算法更快的收敛。图中曲线趋于平缓，即代表着钟摆已经直立，获得奖励为 0。同时，在训练的各个阶段，Dueling-DQN 算法的 cost 值要比 DQN 算法更为接近于 0。综上，Dueling-DQN 算法具有更快的收敛速度及更好的性能。

（下转第 124 页）

5.2 模型的应用

决策树模型准确性高且可解释性好，所以被广泛应用于对业务的透明度和严密性有要求的银行业中。本文的个人信用风险评估模型是使用 Python 和决策树算法构建的，使用训练数据训练好模型之后，模型就可以对用户的个人信用进行风险评估，得出该用户的信用级别是否为好客户，从而预测出该用户的贷款是否违约。基于此评估结果，银行可以明确的解释为什么一个申请者的贷款申请被拒绝或者审批，同时申请者也能知道为什么自己的信用级别不符合银行的要求，且模型以在线模式进行的即时信贷审批可以为银行节约很多人工成本。由此可知，本文的个人信用风险评估模型具有一定的现实意义和价值。

集成学习（ensemble learning）是机器学习里一个重要分支，它体现的思想就是将简单的机器学习模型，也就是弱学习器组合起来，可以得到一个强学习器。本文对于金融诈骗数据的分析预测是基于 XGBoost 模型进行预测的，与传统决策树模型相比较，XGBoost 模型具有对于模型复杂度更好的控制，以及使得学习后的模型不易过拟合的优点，且允许用户在交叉验证时自定义误差衡量方法。

6 结束语

目前，人工智能的方法对于海量数据的处理与分析，以及在运算效率和模型精度方面明显优于传统方法，因此，后续应该加强以及深入对人工智能的学习，推动大数据的发展。与此同时，我国对于信用评估的研究仍然面临着诸多问题与挑战，在个人信用评估领域，单一模型的研究已趋于稳定和成熟，而在组合模型的研究上，还有许多未知模型组合等待人们发掘研究。

参考文献：

[1] 马昕娅 . 我国债券市场信用评级发展现状及问题研究 [J] 时代金融，2018(32):214-216+226.

[2] QUINLAN J R. Induction of decision trees[J]. Machine Learning, 1986, 1(1): 81-106.

[3] QUINLAN J R. C4.5: Programs for Machine Learning[M]. Burlington: Morgan Kaufmanns Publishers, 1993: 69-81.

[4] 张家旺 , 韩光胜 , 张伟 . C5.0 算法在 RoboCup 传球训练中的应用研究 [J]. 计算机仿真 , 2006, 23(4): 132-153.

【作者简介】

何姿娇（1998—），女，湖南岳阳人，本科，专业方向：软件工程大数据技术；

欧阳浩（1979—），男，湖南平江人，硕士，副教授，研究方向：数据挖掘；

刘智琦（1978—），女，广东清远人，硕士，副教授，研究方向：数据挖掘；

付俊宁（1997—），男，广西北海人，本科，专业方向：软件工程 Web 开发；

陈卓婷（2000—），女，广西全州人，本科，专业方向：软件工程大数据技术；

许悦（1999—），女，广西合浦人，本科，专业方向：软件工程后端开发。

（收稿日期：2021-05-06 修回日期：2021-05-29）

（上接第 121 页）

4 结束语

本文对比了传统的 DQN 算法和改变网络结构的 Dueling-DQN 算法，实验表明，在复杂程度不同的环境中，Dueling-DQN 总是要比 DQN 更快的收敛。未来的研究方向会将 Dueling-DQN 算法加入医疗物联网，提升其性能。

参考文献：

[1] SUTTON R, BARTO A. Reinforcement learning: An introduction[M]. Massachusetts: MIT press, 2018.

[2] 肖扬 , 吴家威 , 李鉴学 , 等 . 一种基于深度强化学习的动态路由算法 [J]. 信息通信技术与政策 , 2020(9):48-54.

[3] 李孜恒 , 孟超 . 基于深度强化学习的无线网络资源分配算法 [J]. 通信技术 , 2020, 53(8):1913-1917.

[4] SUN Y, YUAN B, ZHANG T, et al. Research and Implementation of Intelligent Decision Based on a Priori Knowledge and DQN Algorithms in Wargame Environment[J]. Electronics, 2020, 9(10):1668.

[5] BAN T W. An Autonomous Transmission Scheme Using Dueling DQN for D2D Communication Networks[J]. IEEE Transactions on Vehicular Technology, 2020, PP(99):1.

【作者简介】

王恩侃（1997—），男，山西吕梁人，沈阳理工大学信息科学与工程学院电子与通信工程研究生，研究方向：网络与信息安全技术及应用；

姜乐（1996—），女，辽宁朝阳人，沈阳理工大学信息科学与工程学院电子与通信工程研究生，研究方向：现代信号与信息处理技术及应用。

（收稿日期：2021-04-07 修回日期：2021-04-30）