计算机工程与应用
Computer Engineering and Applications
ISSN 1002-8331,CN 11-2127/TP

《计算机工程与应用》网络首发论文

题目：　重复利用状态值的竞争深度 Q 网络算法
作者：　张俊杰，张聪，赵涵捷
网络首发日期：　2020-10-26
引用格式：　张俊杰，张聪，赵涵捷．重复利用状态值的竞争深度 Q 网络算法．计算机工程与应用. https://kns.cnki.net/kcms/detail/11.2127.TP.20201026.1351.016.html

重复利用状态值的竞争深度 Q 网络算法[1]

张俊杰，张　聪*，赵涵捷

武汉轻工大学 数学与计算机学院，武汉 430023

摘　要：在使用反距离加权法(IDW)对土壤重金属含量进行预测时，算法中的超参数一般由先验知识确定，一定程度上存在不确定性。针对这一问题，提出了一种状态值再利用的竞争深度 Q 学习网络算法以精确估计 IDW 的超参数，该算法在训练时，将每轮训练样本中的奖励值进行标准化后，与 Dueling-DQN 中 Q 网络的状态值结合形成新的总奖励值，然后将总奖励值输入到 Q 网络中进行学习，从而增强了状态与动作的内在联系，使算法更加稳定。最后使用该方法在 IDW 上进行超参数学习，并与几种常见强化学习算法进行对比实验。实验表明，提出的 RSV-DuDQN 算法可以使模型更快收敛，同时提升了模型的稳定性，还可以更准确地得到 IDW 的参数估计。

关键词：状态值重利用；竞争深度 Q 学习；反距离加权法；超参数搜索

文献标志码：A　中图分类号：TP391　doi：10.3778/j.issn.1002-8331.2007-0125

张俊杰，张聪，赵涵捷. 重复利用状态值的竞争深度 Q 网络算法.计算机工程与应用
ZHANG Junjie, ZHANG Cong, ZHAO Hanjie. Dueling deep Q network algorithm with state value reuse. Computer Engineering and Applications

Dueling deep Q network algorithm with state value reuse

ZHANG Junjie, ZHANG Cong*, ZHAO Hanjie
Wuhan Polytechnic University School of Mathematics and Computer Science, Wuhan 430023, China

Abstract: When using the inverse distance weighted method (IDW) to predict the content of heavy metals in soil, the super parameters in the algorithm are generally determined by prior knowledge, and there is uncertainty to a certain extent. In order to solve this problem, A competitive deep Q-learning network algorithm for reusing state values is proposed to accurately estimate the hyper parameters of IDW, In the training process, the reward value of each training sample is standardized and combined with the state value of Q network in Dueling-DQN to form a new total reward value, and then the total reward value is input into the Q-network for learning, so as to enhance the internal relationship between state and action and make the algorithm more stable. Finally, this method is used to perform hyperparameter search on the Inverse Distance Weighted method (IDW), and compare experiments with several common deep learning

基金项目：国家自然科学基金面上项目(No.61272278);湖北省重大科技专项资助项目(No.2018ABA099)；湖北省自然科学基金青年项目(No.2018CFB408)；武汉轻工大学引进(培养)人才科研启动项目(No.2019RZ02)。

作者简介：张俊杰（1995-），男，硕士研究生，主要研究方向为人工智能算法优化，计算机视觉，E-mail：1281259317 @qq.com；张聪（1968-），博士，教授，主要研究方向为多媒体信号处理，人工智能算法优化，E-mail：hb_wh_zc@163.com；赵涵捷（1963-），男，博士，教授，主要研究方向为计算机视觉，大数据分析。

algorithms. Experimental results show that the proposed RSV-DuDQN algorithm can make the model converge faster, improve the stability of the model, and get more accurate IDW parameter estimation.
Key words: reuse of state values; dueling DQN; IDW; parameter selection

1. 引言

强化学习（Reinforcement Learning，RL）是一种由动物心理学和控制理论等相关学科结合发展形成的机器学习方法[1-2]。在学习的过程中，强化学习的智能体（agent）通过不断试错的方式进行学习，寻求在当前环境中获得累计奖赏最大的策略[3]。目前强化学习，获得了产业界和科研人员的密切关注，并且在优化、控制、仿真模拟等领域取得了丰富的研究成果[4-6]。深度学习(Deep Learning, DL)是机器学习(Machine Learning, ML)领域中一类重要的方法，其中神经网络是模仿人类大脑的运行机制来解释数据，它可以从人脑无法直接提取特征的复杂高维数据中提取易于区分的特征数据[7]。近年来，深度学习已在计算机视觉，自然语言处理以及语音识别等领域取得较大的进步，也有不少实际应用[8-9]。

在过去几年，强化学习已经和深度学习成功的结合，两者结合形成的机器学习方法称为深度强化学习（Deep Reinforcement Learning，DRL）。如由强化学习中的Q学习（Q-Learning）方法和深度卷积神经网络（Convolutional Neural Networks, CNNs）结合而成的深度Q网络（Deep Q-Network, DQN）是深度强化学习领域中的一个重要方法[10-11]。Van Hasselt等人[12]提出双Q网络（Double DeepQ-Network,DDQN）[13],该方法记在计算目标网络的Q值时使用两套不同的参数，有效的解决了DQN网络对动作值过高的估计，Hausknecht等人[14]首次将长短时间记忆单元（Long-Short Term Memory Neural Network, LSTM）引入DQN中提出了一种基于LSTM修正单元的深度循环Q网络（Deep Recurrent Q-Network, DRQN）,其利用LSTM的记忆功能在大多数Atari2600游戏实验环境中取得较为理想的成绩。Wang等人[15]提出竞争深度Q网络（DuelingDeepQ-Network），将神经网络中提取出来的特征分为优势函数通道和与状态值函数通道输出，该方法显著的提高了在Atari2600环境下的游戏效果。但是使用深度强化学习算法对空间插值算法进行超参数优化时，例如对反距离加权算法中的加权幂次数或克里格插值算法中变异函数模型的基台值、变程等超参数进行优化[16]，当算法的超参数空间大且为连续空间时，优化过程耗时久，效率低并且容易产生过估计现象。

2. 经典理论基础

2.1 反距离加权算法

反距离加权算法广泛应用于重金属含量分析，气象分析，水文分析等多个领域。它是一种多元空间插值方法，通过若干个已知空间离散点的值计算待测点的值。其最大的优点是计算简单且插值速度快。反距离加权算法是根据待测点和已知点的距离的倒数或距离n(n>0)次方的倒数进行加权，然后取所有邻近点的加权平均值。对于点p的估计值Y，其一般形式为[17]：

$$Y(x)=\frac{\sum_{i=1}^{n} w_i(x)Y_i}{\sum_{i=1}^{n} w_i(x)} \quad (1)$$

$$w_i(x)=\frac{1}{d(x,x_i)^p}, \quad (2)$$

$$(i=0,1,\ldots,\ N;\ p\in N)$$

式中：x为插值点；x_i为已知点；Y_i为已知点x_i处的值；N为用于插值的已知点的总数；$d(x, x_i)$为已知点x_i到未知点x的距离。权重w_i随着与未知点距离的增加而减小，p值越大，则距离未知点越近，对未知点的值影响也越大。

2.2 强化学习

强化学习是智能体以当前环境状态为根据，采取行为并从环境中获得奖励的过程，一般情况下，强化学习是以马尔科夫决策过程为基础，寻求马尔科夫决策过程的最佳策略[18-21]。

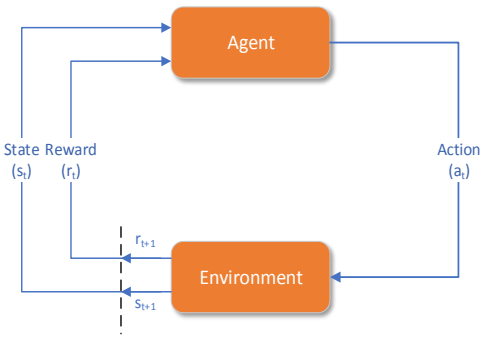

图 1　强化学习框架

Fig.1　Framework of the reinforcement learning

强化学习框架如图 1 所示，在当前状态 S_t 下，总体采取行为 a_t，并根据状态转移函数 P，环境状态将从 S_t 转到 S_{t+1}，同时环境会根据在状态 S_t 下采取行为 a_t 的情况，反馈给智能体一个奖励信号 r。智能体多次循环执行这一过程，以获得最大化累计奖励为目标，通过不断训练，最终得到该过程的最优策略。

2.2.1　DQN 网络

深度 Q 网络是一种经典的深度强化学习算法，其中深度学习部分可以感知环境信息，而强化学习部分可以根据深度学习部分提供的环境信息做出决策，完成从状态到动作的映射，并获得奖赏，再在将这些信息转化为训练数据提供给深度学习，用以持续优化神经网络中的权重矩阵。深度 Q 网络使用神经网络来近似估计 Q-Learning 中的 Q-table 值，但也因此破坏了 Q-Learning 的无条件收敛性[11]。为解决这有问题，DQN 从以下两个方面进行了改进。

第一个方面在 DQN 的智能体与环境的不断迭代交互中，上一个状态与当前状态具有高度相关性，如果不经过处理，直接输入到神经网络中，会导致神经网络产生过拟合现象而无法收敛。因此在 DQN 中加入一个记忆库，用来储存一段时间内的训练样本。在每次学习过程中，DQN 会从记忆库中随机抽取一批样本，输入到深度神经网络中，并对其梯度下降进行学习。在产生新的训练样本时，将老的训练样本和新的训练样本进行混合批次更新，从而在打断相邻训练样本之间的关联性的同时，提高了训练样本的利用率。

第二个方面在 DQN 中建立了一个与当前 Q-evaluate 网络结构完全相同，但参数不同的 Q-Target 的神经网络，该网络仅仅用来计算目标 Q 值，而当前 Q 值只有当前 Q-evaluate 网络预测产生。此方法可以减少目标值与当前值的相关性。损失函数公式为：

$$I = (r + \gamma \max Q_{a'}(s', a'; \omega^{-})) - Q(s, a; \omega))^2 \quad (3)$$

式中，s 表示当前状态，a 表示执行的动作，r 表示环境对智能体的奖励值。$Q(s, a; \omega)$ 为在 s 状态下执行 a 动作时，当前 Q-evaluate 网络的输出值，用来评估当前动态动作对的值函数；$Q(s', a'; \omega^{-})$ 为使用 Q-Target 网络计算得出的目标值函数的 Q 值。

Q-evaluate 网络的参数 w 在每轮训练结束后实时更新，而 Q-Target 网络的参数 w-是由 Q-evaluate 网络的参数 w 延迟更新获得，即在若干轮训练结束后，将 Q-evaluate 网络中的所有参数完整地赋值给 Q-Target 网络。对参数 w 进行求解，可得到值函数的更新公式：

$$\omega_{t+1} = \omega_t + \alpha[r + \gamma \max Q_{a'}(s', a'; \omega^{-}) - Q(s, a; \omega)]\nabla Q(s, a; \omega) \quad (4)$$

2.2.2　Double DQN

在使用经典强化学习算法 Q 学习和深度 Q 学习对动作进行决策和评估时，会参考 Q-max 的值。由于根据 Q-max 选择的动作并非一定是下一状态选择的动作，所以会导致对 Q 现实值的过估计，而为了解决这一问题，Van Hasselt 等人提出了双重深度 q 学习。

DDQN 和经典 DQN 一样也具有两个结构完全相同的神经网络，但 DDQN 与经典 DQN 不同的是：DDQN 通过解耦目标 Q 值动作的选择和目标 Q 值的计算这两步，来消除对 Q 现实值的过度估计问题。先在当前 Q 网络中，找到 Q-max 值对应的动作，然后利用找到的动作在 Q-Target 网络中选择该动作的 Q 值。更新公式为：

$$Y_t = R_{t+1} + \gamma Q(S_{t+1}, \arg\max Q(S_{t+1}, \alpha; \theta_t), \theta_t) \quad (5)$$

3. 基于状态值再利用的竞争深度Q学习网络算法

3.1 竞争深度Q学习网络

在DDQN中，通过减小对目标q值的过估计来优化算法，而在竞争深度Q学习网络（Dueling Deep Q-learning Network, DuDQN）中，通过改进神经网络的结构来优化算法。DuDQN将经典的DQN中的Q网络分成两个部分。第一个部分是价值函数部分，此部分仅仅与状态S有关，和将要采取的动作并无关联，记做$V(s;\theta,\beta)$；第二部分是优势函数部分，此部分不仅与当前状态S有关，并且与将要执行的动作A相关，记做记为$A(s,a;\theta,\alpha)$。所以DuDQN中Q网络的输出为：

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+A(s,a;\theta,\alpha) \quad (6)$$

由于DuDQN的Q网络直接输出Q值，无法分辨价值函数部分和优势函数部分各自的作用，为了体现这种可辨识性，对公式进行适当修改。修改后的公式：

$$Q(s,a;\theta,\alpha,\beta)=V(s;\theta,\beta)+A(s,a;\theta,\alpha)-\max_{a'\in A}A(s,a';\theta,\alpha) \quad (7)$$

在实际应用中，通常使用优势函数的均值来代替优势函数的最大值求解，在保证性能的前提下在一定程度上提高了优化的稳定性。

3.2 RSV-DuDQN模型（Reusing of State Value-Dueling DQN）

在使用DuelingDQN对IDW算法中超参数进行学习时，DuelingDQN的收敛速度和在收敛之后的稳定性方面较其他经典深度强化学习算法有一定差距，算法性能有待提高。针对这一问题，提出了一种改进的DuDQN模型。状态值再利用的竞争深度Q学习网络（ReusingofStateValue - DuelingDeepQ-learningNetwork, RSV-DuDQN）通过将Q网络中的价值函数部分的状态值与当前状态下执行动作的奖励值结合，增强了状态与动作的内在联系，并强化了各个状态-动作对的奖励信号，使得智能体在较好状态时，对环境奖励更加敏感，在较差状态时，对奖励不敏感。从而使算法收敛速度更快，并在收敛之后波动幅度大大减小，提高了算法的稳定性。

在DuDQN训练中，奖励信号值为r，表示在状态s下，执行a动作后，环境对此行为的奖赏。而在RSV-DuDQN中，奖励信号值公式为：

$$R(s,a,p)=r(s,a,p)+\lambda V(s;\theta,\beta) \quad (8)$$

其中：p表示在当前状态s下，执行动作a后，环境转移到下一状态的概率。在对IDW算法中超参数学习中，p是确定的。V(s; w, θ)为Q网络的价值函数部分的输出。λ是惩罚分子，范围为(0, 1]，其作用是确定环境反馈的奖励信号r在整个奖励值中占主导地位，防止因价值函数的状态值过大，导致对环境反馈奖励信号的失去敏感，从而使Q网络无法收敛。

RSV-DuDQN模型的框架流程图如下：

如图2所示，智能体在学习阶段随机采取动作和环境进行交互，并将得到的奖励r与采取的动作a，采取动作前后环境的状态s和s'结合，并以(s,a,r,s')的形式存放到记忆库中。当记忆库中的信息达到规定值后，开始训练。首先从记忆库中随机提取若干条记录输入到DuelingDQN的输入层中，经过若干个隐藏层后到达输出层。图中Fc_V层的输出为价值函数的值$V(s;\theta,\beta)$，而Fc_A层的输出为优势函数的值$A(s,a;\theta,\alpha)$。将Fc_V层的输出与环境的奖励值结合形成最终的总奖励值并将其反馈给智能体。优势函数和价值函数结合形成Q(s,a)，Q(s,a)指导智能体选择动作。如此循环，直到智能体达到目标状态或者训练步数达到指定值。

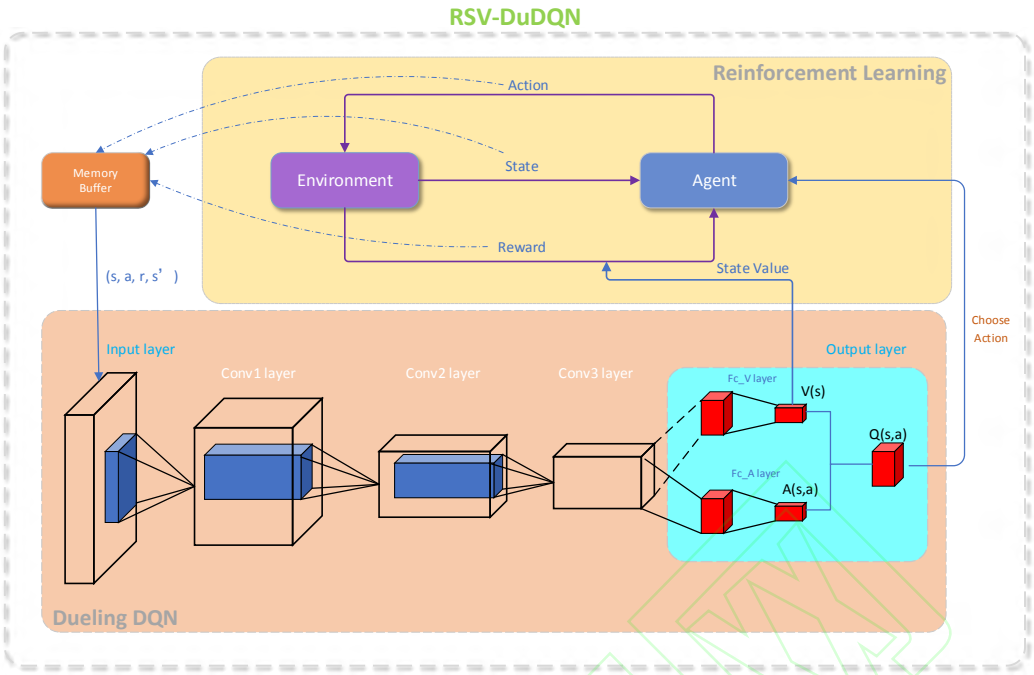

图 2　RSV-DuDQN 框架流程图

Fig.2　Schematic illustration of the RSV-DuDQN

在实际训练中，为防止神经网络学习到奖励信号与状态值的直接关系，使用无限制增加价值函数的状态值来获得更大的环境总奖励。通常在每轮训练时，将从记忆库中抽取的若干个训练样本，将样本中的状态动作对输入到 Q-网络中学习，得到价值函数部分的输出，再将得到的输出进行标准化，然后再以上文提出的方式进行结合。这样做的优点如下：

将不同状态下的训练样本进行标准化，进一步减小了相邻两个状态的样本相关性，更有利于算法的学习和收敛。

标准化之后，得到的状态值仅仅与当前状态有关，切断了总奖励值与 Q 网络的关联。从而避免 Q 网络通过直接输出较大状态值来变相获得总奖励。

算法实现步骤如下：

1) 将经验回放池 D 初始化为容量 N。

2) 用任意权值 θ 初始化动作价值函数 Q 。

3) 用 θ 初始化动作价值函数 $\hat{Q}$。

4) Repeat（对于每个回合）：

5) 初始化序列 $s_1 = \{x_1\}$ ，预处理序列 $\phi_1 = \phi(s_1)$ 。

6) Repeat（对于情节中的每个时间步）：

7) 使用 ε -greedy 算法选择一个动作 a_t 。

否则: $a_t = \arg\max_a Q(\phi(s_t), a; \theta)$ 。

8) 执行动作 a_t ，得到环境的奖赏值 r_t 和当前环境的观测值 x_{t+1} 。

9) 设置 $s_{t+1} = s_t$ ，并预处理图像 $\phi_{t+1} = \phi(s_{t+1})$ 。

10) 将 $(\phi_t, a_t, r_t, \phi_{t+1})$ 储存到经验回放池 D 中。

11) 从 D 中抽取 minibatch 个样本 $(\phi_j, a_j, r_j, \phi_{j+1})$ 。

12) 标准化取出的minibatch个样本中的 r_j 得到 $r_j^{'}$ 。

13) 将 (ϕ_j, a_j) 输入到 Q-估值网络得到状态值 SV_j 。

14) 设置 $R_j = r_j^{'} + \lambda * SV_j$ 。

15) 如果在 *j+1* 步回合结束，则： $y_j = R_j$ 。

否则： $y_j = R_j + \gamma \max_{a'} \hat{Q}(\phi_{j+1}, a'; \theta^-)$ 。

16) 对损失函数 $L = (y_j - Q(\phi_j, a_j; \theta))^2$ 中的参数 θ 执行梯度下降步骤。

17) 每 C 步将 $\hat{Q}$ 重置为 Q 。

18) Unitil 智能体达到当前情节终止状态。

19) Unitil 达到预期训练次数。

4. 实验与结果分析

为验证提出的 RSV-DuDQN 模型较常见的深度强化学习模型具有一定的优势。实验一分别使用 DQN，DDQN，DuDQN 和 RSV-DuDQN 学习反距离加权插值法在武汉城郊农田土壤重金属含量数据集[22]上的加权幂次数，该数据集来自于湖北省技术创新重大项目“武汉城郊农田土壤重金属积累特征及风险评价”，数据集中每个样本的测定方法都是依据《土壤环境监测技术规范》（HJ/T166-2004）和《土壤环境质量 农业农用地土壤环境污染管控》（GB15618-2018）的要求执行，总采样点 1161 个，重金属种类 8 种。实验二使用由实验一学习到的超参数进行反距离加权插值并与经典反距离加权插值算法进行对比。

实验环境如下：处理器为 AMD2600，主频为 3.4GHz，内存为 24G，由于模型中使用深度神经网络，大多采用矩阵运算，因此使用了 GTX1660 图形处理器对模型进行辅助加速运算。

4.1 基于 RSV-DUDQN 的超参数估计

为验证 RSV-DUDQN 模型的有效性，本部分采用武汉城郊农田土壤重金属含量数据集，该数据集包括 As，Cd，Cr，Cu，Hg，Ni，Pb，Zn 八种常见土壤重金属。分别使用 DQN，DDQN，DuDQN 和 RSV-DuDQN 估计反距离加权插值法在该数据集中六种金属含量数据上的超参数。所有深度强化学习算法中智能体的动作空间为[-1, 1]，并且经过多次实验，最终确定动作空间离散为[-1, -0.5, -0.1, 0, 0.1, 0.5, 1]。经过实验验证，将动作由连续空间，离散到精度 0.1 的离散空间后，算法学习到的超参数对整个插值结果影响可以忽略不记。在实验开始阶段，先使用 ArcGIS+软件将原数据中的经纬网坐标转换为常用的平面直角坐标，并其标准化与初始化后的超参数一起输入到 Q 网络。八种金属的算法训练图如图 3~图 10 所示，横坐标为训练次数，单位为(次)，纵坐标代表在当前学习到的超参数下，用反距离加权法进行插值得到预测值与真实值的误差，单位为(mg/kg)。四种深度强化学习算法分别在对八种重金属含量进行 IDW 的超参数预测时，训练情况如表 1。表中展示了各种模型在对不同重金属数据集训练中第一次收敛时的训练轮数，单位为（轮）。为了更直观地展示算法训练时的情况，在训练中，当训练轮数达到 5000 时，停止训练，此时还未收敛的算法其在表格中的收敛时的训练次数以“>5000”形式表达。

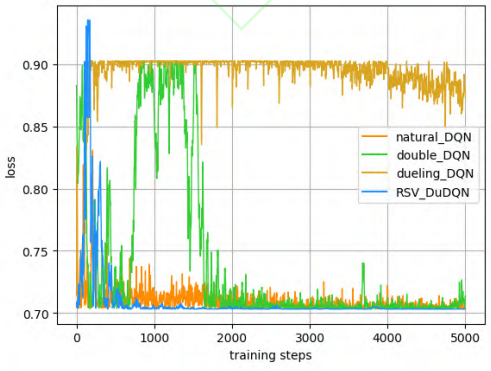

图 3　重金属 As 数据的训练结果

Fig.3　Training on dataset of As

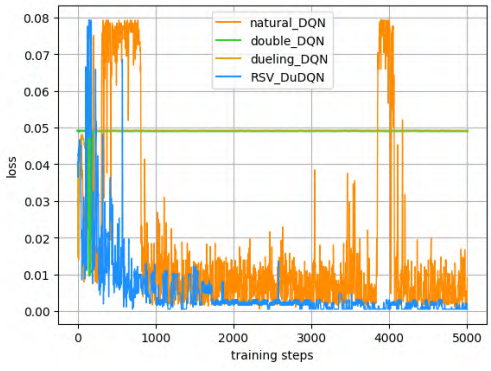

图 4　重金属 Cd 数据的训练结果

Fig.4　Training on dataset of Cd

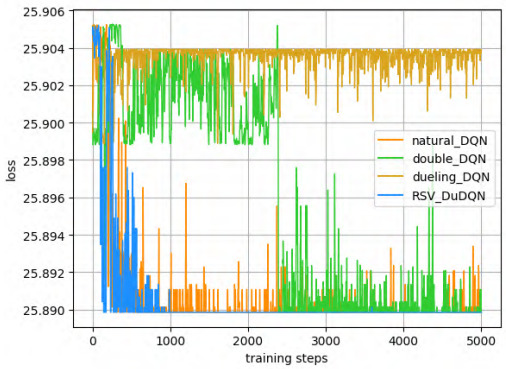

图 5　重金属 Cr 数据的训练结果

Fig.5　Training on dataset of Cr

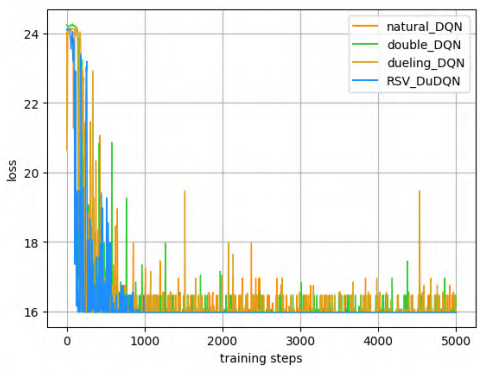

图 6　重金属 Cu 数据的训练结果

Fig.6　Training on dataset of Cu

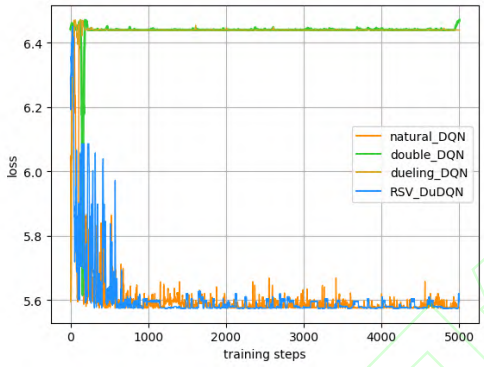

图 7　重金属 Ni 数据的训练结果

Fig.7　Training on dataset of Ni

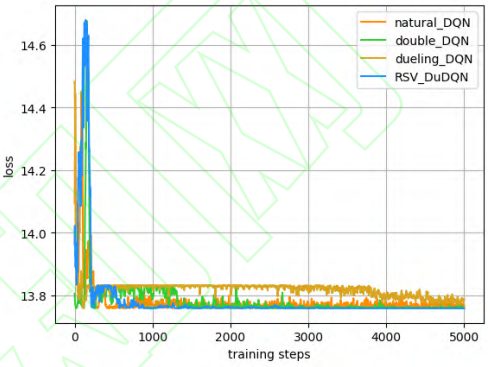

图 8　重金属 Pb 数据的训练结果

Fig.8　Training on dataset of Pb

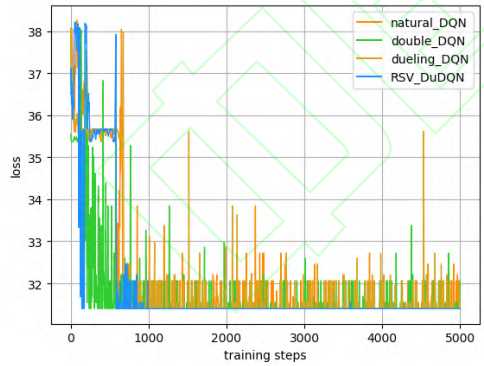

图 9　重金属 Zn 数据的训练结果

Fig.9　Training on dataset of Zn

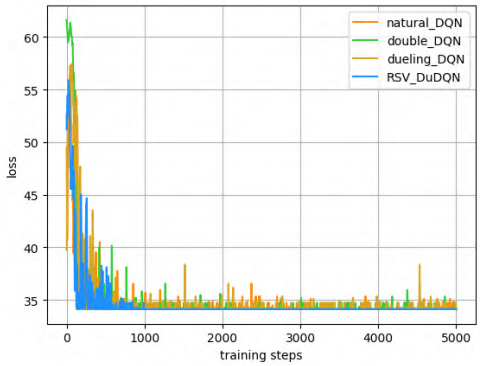

图 10　重金属 Hg 数据的训练结果

Fig.10　Training on dataset of Hg

表 1　不同模型在不同数据集上的最小收敛轮数

Table 1　Minimum convergence steps of training on different data sets with different models

	As	Cd	Cr	Cu	Ni	Pb	Zn	Hg
DQN	2856	1306	1961	161	183	847	683	1309
DDQN	600	>5000	1098	122	3824	>5000	603	3249
DuDQN	>5000	>5000	4875	200	>5000	>5000	900	>5000
RSV-DuDQN	468	1011	956	118	138	609	603	716

表 2　模型在不同数据上的插值误差

Table 2　Loss obtained by interpolation with different models

	Model	MSE	RMSE	MAPE(%)	MAE
id=0	Common parameter	26.79	5.18	29.24	3.36
id=0	RSV-DuDQN	16.17	4.02	19.84	2.31
id=1	Common parameter	165.42	12.85	74.15	9.68
id=1	RSV-DuDQN	111.68	10.57	53.44	6.67
id=2	Common parameter	123.99	11.14	31.67	8.09
id=2	RSV-DuDQN	76.25	8.73	22.44	5.71

由于 RSV-DuDQN 的时间复杂度与 DQN，DDQN 和 DuDQN 相同，因此各模型的最小收敛轮数基本可以代表各模型的收敛时间。由表 1 可知，对于不同的重金属种类，DQN，DDQN 和 DuDQN 的算法收敛速度不同。其中 DDQN 在 As 的数据上收敛较快，DuDQN 在 Cu 的数据上收敛较快，DQN 在 Cd，Cr，Ni 数据上的收敛速度相比于 DDQN 和 DuDQN 的收敛速度有较大提升。在 As，Cd，Cr，Ni，Pb 的数据上，RSV-DQN 模型在收敛速度方面明显优于其他三个模型，而在 Hg 数据上，四个模型的收敛速度相同。由图 5，图 6，图 7，图 10 可以看出，DQN，DDQN 和 DuDQN 在算法搜寻到最优解之后，仍然会有出现较大波动，无法稳定在较好的状态，此情况在图 5 中 DQN 模型的表现上尤为明显。而由图 4，图 5，图 9 可知，DDQN 和 DuDQN 并不能总是学习到最优超参数，某些情况下仅仅可以学习到较优超参数。对于图 3-图 10，RSV-DuDQN 模型总是可以较快找到最优解并且可以一直稳定在一定范围内，说明该模型相比于其他模型，在稳定性方面具有一定的优越性。

4.2 基于 RSV-DuDQN 的 IDW 插值实验

为了验证由深度强化学习模型搜索出来的超参数的有效性，本次实验使用实验一中数据集的江夏区数据，共包含 266 个采样点。分别使用 RSV-DuDQN 模型搜索出来的超参数和常用先验超参数进行 IDW 插值实验，在 As 数据集上使用 RSV-DuDQN 模型搜索出来的超参数进行 IDW 插值实验，然后使用相同方法常见超参数进行对比实验，并标记为“id=0”。用相同方法在 Cr 和 Ni 数据上进行实验，分别标记为“id=1”和“id=2”。最后两个模型预测的结果与真实值作比较，得到均方误差（Mean Square Error，MSE）、均方根误差（Root Mean Square Error，RMSE）、平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）和平均绝对误差（Mean Absolute Error，MAE），实验结果如表 2 所示，所有误差精度取 0.01。

在三次对比实验中，由于加入了 RSV-DuDQN 模型，整个插值过程更为复杂，除 IDW 插值所需时间外，RSV-DuDQN 模型需要额外的

时间对 IDW 算法中的超参数进行学习。，但由表 2 可知，基于 RSV-DuDQN 的反距离加权法的 MSE，RMSE，MAPE 以及 MAE 均在不同程度小于经典反距离加权法，使用 RSV-DuDQN 模型搜索出来的超参数进行 IDW 插值时得到的平均误差相比与使用常见超参数插值时的误差低 13.11%，说明其在该数据集上的插值表现优于经典反距离加权法，因此可知，由 RSV-DuDQN 模型学习到的超参数优于常见的先验超参数，并且，RSV-DuDQN 模型确实有效可行。

5. 结束语

本文提出了一种基于竞争深度 Q 学习网络的 RSV-DuDQN 模型。该模型将竞争深度 Q 学习算法中 Q 网络中价值函数部分的状态输出值与环境反馈得到的奖励信号相结合，并将其以总奖励的形式加入到强化学习的训练中。解决了竞争深度 Q 学习网络在一定情况下收敛速度较慢并且收敛之后网络依然不稳定的问题。在与 DQN，DDQN 和 DuDQN 的对比实验中，证明了 RSV-DuDQN 模型在收敛速度以及稳定性确实具有一定的优势。最后使用 RSV-DuDQN 模型学习到的超参数对数据进行反距离加权法插值，并与常用先验超参数进行对比实验，实验证明了 RSV-DuDQN 模型学习到的超参数具有一定可行性。

虽然该算法在插值算法类的小规模动作空间中的超参数优化问题上，优化效果较好，但在较大规模动作空间上超参数优化过程的时间复杂度还有进一步提高空间，在接下来的研究中，可以使用遗传算法对深度强化学习的超参数进行优化，进一步在降低算法时间复杂度的同时提高算法收敛速度。

参考文献：

[1] Sutton, R, Barto, A. Reinforcement Learning:An Introduction[M]. MIT Press, 1998.

[2] Lee D ,Seo H , Jung M W . Neural basis of reinforcement learning and decision making.[J]. Annual Review of Neuroence, 2012, 35(1):287.

[3]黄东晋,蒋晨凤,韩凯丽.基于深度强化学习的三维路径规划算法[J/OL].计算机工程与应用, 2020, 56(15): 30-36.
(HUANG D J, JIANG C F, HAN K L. 3D path planning algorithm based on deep reinforcement learning[J/OL]. Computer Engineering and Applications, 2020, 56(15):30-36)

[4] Xu X, Zuo L, Huang Z. Reinforcement learning algorithms with function approximation: Recent advances and applications[J]. Information ences, 2014, 261:1-31.

[5] Liu Q, Zhang X, Zhu F, et al. Protein interaction network constructing based on text mining and reinforcement learning with application to prostate cancer[J]. Systems Biology Iet, 2015, 9(4):106-112.

[6] David Silver, Richard S. Sutton, Martin Müller. Temporal-difference search in computer Go[J]. Machine Learning, 2012, 87(2): 183-219.

[7] Lecun Y, Bengio Y, Hinton G. Deep Learning[J]. Nature, 2015, 521(7553): 436-444.

[8] Zhang Q , Yang L T , Chen Z , et al. A survey on deep learning for big data[J]. Information Fusion, 2018, 42:146-157.

[9] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.

[10] Li H, Wei T, Ren A, et al. Deep reinforcement learning: Framework, applications, and embedded implementations[C]//2017 IEEE/ACM International Conference on Computer-Aided Design (ICCAD). IEEE, 2017: 847-854.

[11] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[12] Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Phoenix, USA, 2016: 2094-2100.

[13] Van Hasselt H. Double Q-learning[J]. Advances in Neural Information Processing Systems, 2010, 2613-2621.

[14] Hausknecht M, Stone P. Deep recurrent q-learning for partially observable mdps[J]. arXiv preprint arXiv:1507.06527, 2015.

[15] Wang Z, Freitas N D, Lanctot M. Dueling network architectures for deep reinforcement learning[C]//Proceedings of the 32nd International Conference on Machine Learning. New York, USA, 2016, 529(7587): 484-489

[16] Liu Y, Zhang C. Application of Dueling DQN and DECGA for Parameter Estimation in Variogram Models[J]. IEEE Access, 2020, 8:38112-38122.

[17] Renka R J. Multiriate interpolation of large sets of scattered data[J]. ACM Transactions on Mathematical Software, 1988, 14(2): 139-148

[18] Wu Y, Shen T. Polivy Iteration algorithm for optimal control of stochastic lagical dynamical systems[J]. IEEE Transactions on Neural Networks & Learning Systems, 2017, 28(99): 1-6

[19] Wei Q, Liu D, Lin H. Value iteration adaptive dynamic programming for optimal control of discrete-time nonlinear systems[J]. IEEE Transactions on Cybernetics, 2018, 46(3): 840-853.

[20] ZHU F, WU W, FU Y C, et al. Security depth reinforcement learning method based on double depth network[J]. Journal of Computer Science, 2019, 42(8): 1812-1826.

（朱斐,吴文,伏玉琛,刘全.基于双深度网络的安全深度强化学习方法[J].计算机学报，2019，42(8): 1812-1826.）

[21] CHEN J P, ZOU F, LIU Q, et al. reinforcement learning algorithm based on generative adversary network[J],Computer Science, ,2019,46(10):265-272.

(陈建平,邹锋,刘全,吴宏杰,胡伏原,傅启明.一种基于生成对抗网络的强化学习算法[J].计算机科学, 2019, 46(10):265-272.)

[22] Zhang M X , Li H . Dataset of Heavy Metals in Surface Soil of Yinchuan City, Ningxia Hui Autonomous Region, China[J]. Journal of Global Change Data & Discovery, 2018, 2(2):198-204.