软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2021,32(11):3496−3511 [doi: 10.13328/j.cnki.jos.006077]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

一种基于广义异步值迭代的规划网络模型*

陈子璇[1], 章宗长[1], 潘致远[2], 张琳婧[2]

[1](计算机软件新技术国家重点实验室(南京大学),江苏 南京 210023)

[2](苏州大学 计算机科学与技术学院,江苏 苏州 215006)

通讯作者: 章宗长, E-mail: zzzhang@nju.edu.cn

摘 要: 近年来,如何生成具有泛化能力的策略已成为深度强化学习领域的热点问题之一,并涌现出了许多相关的研究成果,其中的一个代表性工作为广义值迭代网络.广义值迭代网络是一种可作用于非规则图形的规划网络模型.它利用一种特殊的图形卷积算子来近似地表示状态转移矩阵,使得其在学习到非规则图形的结构信息后,可通过值迭代过程进行规划,从而在具有非规则图形结构的任务中产生具有泛化能力的策略.然而,由于没有考虑根据状态重要性来合理分配规划时间,广义值迭代网络中的每一轮迭代都需要在整个状态空间的所有状态上同步执行.当状态空间较大时,这样的同步更新会降低网络的规划性能.用异步更新的思想来进一步研究广义值迭代网络.通过在值迭代过程中定义状态优先级并执行异步值更新,提出了一种新型的异步规划网络模型——广义异步值迭代网络.在未知的非规则结构任务中,与广义值迭代网络相比,广义异步值迭代网络具有更高效且更有效的规划过程.进一步地,改进了广义值迭代网络中的强化学习算法及图形卷积算子,并通过在非规则图形和真实地图中的路径规划实验验证了改进方法的有效性.

关键词: 深度学习;强化学习;模仿学习;规划;异步更新

中图法分类号: TP181

中文引用格式: 陈子璇,章宗长,潘致远,张琳婧.一种基于广义异步值迭代的规划网络模型.软件学报,2021,32(11):3496–3511. http://www.jos.org.cn/1000-9825/6077.htm

英文引用格式: Chen ZX, Zhang ZZ, Pan ZY, Zhang LJ. Planning network model based on generalized asynchronous value iteration. Ruan Jian Xue Bao/Journal of Software, 2021,32(11):3496−3511 (in Chinese). http://www.jos.org.cn/1000-9825/6077.htm

Planning Network Model Based on Generalized Asynchronous Value Iteration

CHEN Zi-Xuan[1], ZHANG Zong-Zhang[1], PAN Zhi-Yuan[2], ZHANG Lin-Jing[2]

[1](State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China)

[2](School of Computer Science and Technology, Soochow University, Suzhou 215006, China)

Abstract: In recent years, how to generate policies with generalization abilities has become one of the hot issues in the field of deep reinforcement learning, and many related research achievements have appeared. One representative work among them is generalized value iteration network (GVIN). GVIN is a differential planning network that uses a special graph convolution operator to approximately represent a state-transition matrix, and uses the value iteration (VI) process to perform planning during the learning of structure information in irregular graphs, resulting in policies with generalization abilities. In GVIN, each round of VI involves performing value updates synchronously at all states over the entire state space. Since there is no consideration about how to rationally allocate the planning time according to the importance of states, synchronous updates may degrade the planning performance of network when the state space is

* 基金项目: 国家自然科学基金(61876119); 江苏省自然科学基金(BK20181432); 中央高校基本科研业务费专项资金(022114380010)

Foundation item: National Natural Science Foundation of China (61876119); Natural Science Foundation of Jiangsu Province (BK20181432); Fundamental Research Funds for the Central Universities (022114380010)

收稿时间: 2019-11-12; 修改时间: 2020-03-17; 采用时间: 2020-04-30

large. This work applies the idea of asynchronous update to further study GVIN. By defining the priority of each state and performing asynchronous VI, a planning network is proposed, it is called generalized asynchronous value iteration network (GAVIN). In unknown tasks with irregular graph structure, compared with GVIN, GAVIN has a more efficient and effective planning process. Furthermore, this work improves the reinforcement learning algorithm and the graph convolutional operator in GVIN, and their effectiveness are verified by path planning experiments in irregular graphs and real maps.

Key words: deep learning; reinforcement learning; imitation learning; planning; asynchronous update

近几年,随着深度学习在人工智能领域的流行,神经网络模型已被广泛应用于强化学习(reinforcement learning,简称 RL)和模仿学习(imitation learning,简称 IL)等机器学习任务中,并取得了很多成果[1−8].在这些任务的解决方案中,策略通常用神经网络来表示.然而,由于网络中缺少明确的规划模块和相应的规划运算,这种网络形式的策略本质上是反应式的[9].由于反应式策略无法理解动作的目标导向性,因此采用这种策略的智能体(agent)通常只能学会解决在训练集中出现过的任务,而较难泛化到解决其训练集之外的未知任务[10],从而在实际应用中会遇到很大的挑战.

为了解决这个挑战,Tamar 等人[10]提出了一种嵌有值迭代模块的可微的规划网络——值迭代网络(value iteration network,简称 VIN).该网络可利用 IL 或 RL 算法进行端到端的训练,使得网络在未知任务中能执行规划运算,从而生成具有较好泛化能力的策略.VIN 中,值迭代模块的关键创新之处在于:它以一种堆叠式的卷积神经网络[9]来模拟值迭代过程[11],使得智能体可以顺利学习到当前任务中的动态信息,进而利用规划方法得到有效的且具有泛化能力的策略.然而,由于其值迭代模块中的卷积算子在内部结构上具有局限性,目前 VIN 的应用领域仅限于具有规则结构的任务,即内部构成为一维顺序结构或是二维栅格结构的任务.在自动驾驶汽车的路径规划、网页中信息采集/导航等内部构成为非规则结构的任务中,智能体会无法准确地学习到非规则环境的动态信息,从而无法进行有效的规划.因此,Niu 等人[12]提出了一种基于 VIN 的广义值迭代网络(generalized value iteration network,简称 GVIN)来消除这种局限性.GVIN 通过两个方面改进了 VIN:(1) 它利用一种适用于非规则图形的图形卷积算子来近似表示状态转移矩阵,以模拟值迭代过程.该卷积算子泛化了 VIN 中所使用的二维图形卷积算子,使得其能够不受规则图形结构的限制,从而作用于具有非规则结构的任务中.(2) 它提出了一种 n 步 Q 学习算法[13]的改进算法——情节式 Q 学习算法(episodic Q-learning),使得规划网络在利用 RL 算法训练时的稳定性有了进一步的提升.由于 GVIN 成功地将 VIN 的应用范围扩大至具有非规则图形结构的任务中,所以称它为“广义的(generalized)”.

然而,VIN 和 GVIN 中均存在着一个相同的问题——这两个网络中所模拟的值迭代过程均为同步执行的,即无论每个状态的重要性如何,整个状态空间中所有状态的值函数在每一轮值迭代过程中都会被更新.这意味着网络并没有根据状态的重要性来合理分配每个状态所需的规划时间,那么当状态空间较大时,规划过程可能会长时间陷入无意义的值更新中,导致网络整体规划性能的下降[14].

基于这两个规划网络中应用范围更为广泛的 GVIN 模型,本文提出了一种改进的异步规划网络模型,即广义异步值迭代网络(generalized asynchronous value iteration network,简称 GAVIN).为了实现 GAVIN,本文依据异步更新[15,16]的思想,提出了一种适用于 GVIN 的异步更新方法——基于状态的异步更新方法,并将其进一步地应用于 GVIN 的值迭代过程中.该方法的主要思想是:在每轮值迭代过程开始之前,为状态空间上的每个状态定义其优先级,其后根据状态优先级来异步更新状态值,即使得状态空间上某些状态处的值被更新之前,那些在规划过程中相对更为重要的状态的值已被多次更新,从而合理地分配规划过程中智能体在每个状态上所需的规划时间.需要指出的是:文献[15,16]中所提出的异步更新方法仅适用于具有规则结构的规划任务,而基于状态的异步更新方法不仅适用于具有规则结构的任务,还能更好地应用于求解具有非规则结构的任务.此外,GAVIN 中的异步更新过程会根据当前环境的变化来自适应地选择需要更新的状态集合,且该集合的大小并非为固定值,这也与文献[15,16]中的方法有所不同.

与 GVIN 相同,GAVIN 使得智能体能够在具有非规则图形结构的未知任务中自我学习环境的动态信息并规划出最优策略.此外,通过使用基于状态的异步更新方法,GAVIN 有效地解决了原网络模型规划过程中存在的

规划时间分配不合理的问题,进一步避免了无意义的值更新过程,提高了其在具有非规则结构的任务中的规划效率及泛化能力.这个改进的规划网络模型能为许多实际应用场景带来益处.例如:它可被应用于自动驾驶领域中,使得自动驾驶汽车在未知路况中的路径规划过程更为高效且有效.值得注意的是:GAVIN中的规划算法与传统的规划算法不同,如 Dijkstra 算法[17],后者在规划过程中需要一个已知的环境模型,而前者旨在通过试错(trial-and-error)或模仿专家样本的数据来学习一个广义的环境模型,使训练后的网络模型能应用于与训练任务不同的任意未知任务中.其次,为了进一步提高规划网络中 RL 算法的训练性能,本文将加权双 Q 学习(weighted double Q-learning)[18]中所用的加权双估计器(weighted double estimator)思想与情节式 Q 学习相结合,提出了一种新的 RL 训练算法——情节式加权双 Q 学习(episodic weighted double Q-learning).最后,本文提出一种新的定义方法来小幅改进 GVIN 中所用的、由基于嵌入信息的核函数所定义的图形卷积算子[12],使得利用这个改进后的卷积算子的网络在规划过程中能够更为准确地学习到非规则图形的基本结构信息,从而获得更好的规划性能及泛化能力.

本文的具体实验场景为智能体在非规则图形及真实路况地图中的路径规划问题.在真实路况地图环境中,每个路口可被形式化为非规则图形中的节点,每条道路可被形式化为非规则图形中的边.在这些实验场景中,每个节点都具有不同的局部结构,即每个节点所连接的节点数目不同且相连节点之间边的方向也不同.使用具有非规则图形结构的实验环境验证了 GAVIN 的广义性.实验结果有力地验证了新方法的有效性.与 GVIN 相比,在利用内部组成结构较为简单的非规则任务训练过后,GAVIN 所表示的策略能够在更复杂且更大规模的未知测试任务中获得更好的泛化性能.具体地,本文分别利用美国明尼苏达州高速地图(Minnesota highway map)以及纽约市区街道地图(New York city street map)的真实数据对新方法进行评估,实验结果有力地验证了 GAVIN 在大规模实际应用场景中的适用性和有效性.

1 基础知识及相关工作

本节对本文内容所涉及的基础知识及相关工作进行了介绍.第 1.1 节中介绍了马尔可夫决策过程,第 1.2 节对 GVIN 模型进行了简要介绍,第 1.3 节介绍了相关的 RL 算法——情节式 Q 学习算法及加权双 Q 学习算法.

1.1 马尔可夫决策过程

许多序贯决策问题都可以用马尔可夫决策过程(Markov decision process,简称 MDP)[19]来建模.MDP 可表示为一个五元组(S,A,Tr,R,γ),其中,S 是状态空间,A 是动作空间,$Tr(s'|s,a)$是状态转换函数,$R(s,a)$是奖赏函数,$\gamma\in(0,1)$是折扣因子.MDP 中的策略π是指从状态空间 S 到动作空间 A 的映射.在策略π下,状态 s 的值为 $V^{\pi}(s)=\mathbb{E}^{\pi}[\sum_{t=0}^{\infty}\gamma^{t}R(s_t,a_t)\,|\,s_0=s]$. 在$\pi$下,状态-动作对$(s,a)$的值为 $Q^{\pi}(s,a)=\mathbb{E}^{\pi}[\sum_{t=0}^{\infty}\gamma^{t}R(s_t,a_t)\,|\,s_0=s,a_0=a]$.智能体求解 MDP 的目标为:找到最优策略π^*,以最大化其期望回报.当 MDP 模型已知时,最优策略可以通过值迭代过程来获得.值迭代过程中包含两个子过程:$V_{n+1}(s)=\max_a Q_n(s,a)$, $Q_n(s,a)=R(s,a)+\gamma\sum_{s'}Tr(s'\,|\,s,a)V_n(s')$. 通过这两个过程,随着 $n\to\infty$,Q_n 可渐近收敛到最优值 Q^*.由此可得最优策略$\pi^*(s)=\mathrm{argmax}_a Q^*(s,a)$.

在 RL 问题中,智能体的目的是通过与环境交互,从环境给予的奖赏信号中学习到一个最优策略.即智能体在未知的环境中,通过不断的试错来进行学习,以找到能够最大化期望累积奖赏的策略[19].

在IL问题中,智能体的学习过程有所不同,它不是从环境提供的奖赏信号中学习,而是从专家提供的演示数据中学习.即智能体从一组专家样本中学习其要执行的策略.一般而言,每一个专家样本均包含了一种具体情况的详细描述以及在这种情况下,智能体应采取的正确动作的规范(标签)[3,19].

1.2 广义值迭代网络(GVIN)模型

GVIN 是一种嵌有规划模块的可微的规划网络模型,利用这个规划模块,GVIN 能够学习到非规则图形中的环境动态信息,并利用这些信息进行规划,最终生成具有泛化能力的策略.图 1 为 GVIN 的整体网络结构示意图.图中左上角为输入到网络进行训练的 8-节点的非规则图形 G.f_R 为用于生成图形内部各节点奖赏信号的恒等函数,该函数的输入信息为经过图形信号{0,1}编码后的非规则图形,其中,只有目标节点的信号值为 1,其他节点的

信号值均为 0.f_P 为用于生成图形卷积算子的函数,其中,训练参数 $\boldsymbol{w_P}$ 用于参数化图形卷积算子.函数 f_P 的输入信息及内部具体结构将在第 1.2.1 节中进行介绍.$\boldsymbol{R}$,$\boldsymbol{P}$,$\boldsymbol{V}$,$\boldsymbol{Q}$ 分别表示非规则图形的奖赏信号、图形卷积算子、状态值图形信号以及状态-动作值图形信号.由于非规则图形的内部结构特性,这 4 个信号值在 GVIN 规划模块的计算过程中均以矩阵向量的形式表示.

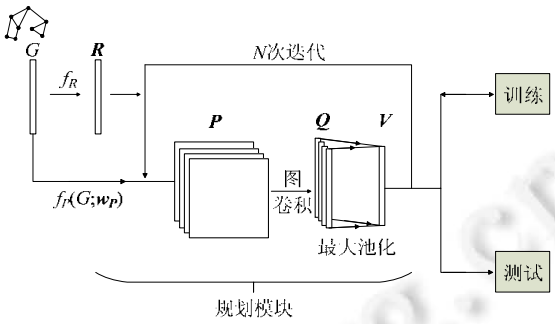

Fig.1 Overall architecture of GVIN

图 1 GVIN 的整体结构

GVIN 规划模块中的第 n 轮值迭代过程可被形式化为:

$$\boldsymbol{Q}_{n+1}^{(a)} = \boldsymbol{P}^{(a)}(\boldsymbol{R} + \gamma \boldsymbol{V}_n) \tag{1}$$

$$\boldsymbol{V}_{n+1} = \max_a \boldsymbol{Q}_{n+1}^{(a)} \tag{2}$$

在 GVIN 的网络结构中,公式(1)以卷积层的形式呈现,图形卷积算子 $\boldsymbol{P}$ 相当于卷积核,其上的每个通道对应于智能体的每个动作,$\boldsymbol{P}^{(a)}$表示第 a 个通道上的图形卷积算子.公式(2)以最大池化层的形式呈现.

GVIN 中的规划模块近似地模拟了值迭代的过程.在 N 次迭代后,网络会获得图中各节点(即 MDP 中的状态)的值函数,并最终利用这些值函数进行策略规划.迭代次数 N 的值,根据输入图形的大小及训练算法的种类来设置.GVIN 中的网络参数利用 IL 或 RL 算法进行训练.在 RL 算法的训练过程中,智能体采取ε贪心策略选择动作.在测试过程中,智能体采取贪心策略选择动作.

1.2.1 基于嵌入信息的核函数

从数学定义上来说,一个加权无向图可以被表示为 $G=(v,\boldsymbol{X},E)$的形式,其中,$v=\{v_1,\ldots,v_N\}$表示一组节点;$\boldsymbol{X}$ 指节点嵌入信息,第 i 个节点的嵌入信息为 $\boldsymbol{X}_i$;E 表示一组边.如果图形的节点数目为 n,那么每个图形都可以用大小为 $n\times n$ 的邻接矩阵 $\boldsymbol{A}$ 来表示.如果 v_i 和 v_j 之间有边相连接,则 $\boldsymbol{A}_{i,j}$=1;否则,$\boldsymbol{A}_{i,j}$=0.在 GVIN 中,用于进行非规则图形图卷积操作的图形卷积算子被形式化为 $\boldsymbol{P}=f_P(G;\boldsymbol{w_P})$,其中,每个元素的基本定义为 $\boldsymbol{P}_{i,j} = \boldsymbol{A}_{i,j} K_{\boldsymbol{w_P}}(\boldsymbol{X}_i, \boldsymbol{X}_j)$,其中,核函数 $K_{\boldsymbol{w_P}}(\cdot,\cdot)$ 由 $\boldsymbol{w_P}$ 进行参数化.这个定义意味着:无论是根据哪种核函数定义,输入 GVIN 的每张非规则图形的图形卷积算子都是由其邻接矩阵和特定的核函数来共同定义的.GVIN 共提出了 3 种用于定义图形卷积算子的核函数,本文只介绍其中能使得网络具有最优泛化能力的那一个——基于嵌入信息的核函数(embedding-based kernel).

通过使用基于嵌入信息的核函数来定义的图形卷积算子,GVIN 能准确地获取非规则图形中隐藏的结构信息,从而能在整张图形的每个节点上进行规划.在使用基于嵌入信息的核函数进行定义的图形卷积算子中,(i,j)节点之间转移概率的定义为(GVIN 原文中该公式的定义有误,本文中该公式的定义已被修正)

$$\boldsymbol{P}_{i,j} = \frac{I_{i=j} + \boldsymbol{A}_{i,j}}{\sqrt{\sum_k (I_{k=j} + \boldsymbol{A}_{k,j}) \sum_k (I_{k=i} + \boldsymbol{A}_{i,k})}} K_{emb}(\boldsymbol{X}_i, \boldsymbol{X}_j) \tag{3}$$

当 $i=j$ 时,指示函数 $I_{i=j}$=1;否则为 0.$\boldsymbol{A}$ 为图形的邻接矩阵,若 i,j 节点之间有边相连,则 $\boldsymbol{A}_{i,j}$=1;否则,$\boldsymbol{A}_{i,j}$=0.基于嵌入信息的核函数为 $K_{emb}(\boldsymbol{X}_i,\boldsymbol{X}_j)=mnnet([\boldsymbol{X}_i-\boldsymbol{X}_j])$,其中,$mnnet(\cdot)$表示一个标准的多层神经网络,$\boldsymbol{w_P}$ 为该网络中的权重.

$\frac{I_{i=j}+\boldsymbol{A}_{i,j}}{\sqrt{\sum_k(I_{k=j}+\boldsymbol{A}_{k,j})\sum_k(I_{k=i}+\boldsymbol{A}_{i,k})}}$为激活系数,该系数利用图形邻接矩阵中潜在的节点连接性来激活核函数.

1.3 相关RL算法

1.3.1 情节式 Q 学习

情节式 Q 学习[12]是 n 步 Q 学习的一种改进算法.当这两个算法与神经网络模型相结合时,它们的区别在于: n 步 Q 学习算法使用两个结构相同的网络模型来共同训练网络参数,即目标网络和行为网络.算法中每个情节的持续时间固定,每隔 n 步后,计算每一步的损失函数及梯度,累计梯度,并以此更新网络参数.n 步中每一时间步的损失函数为 $(G_{\bar{t}+i}-Q(s_{\bar{t}+i},a;\theta))^2$,其中,$G_{\bar{t}+i}\leftarrow R_{\bar{t}+i+1}+\gamma G_{\bar{t}+i+1}$,$i\in\{0,1,2,\ldots,n-1\}$.$G$ 为累积奖赏,其初始值为

$$G_{\bar{t}+n}\leftarrow\begin{cases}0, & s_{\bar{t}+n}\text{为目标状态}\\ \max_{a'}Q(s_{\bar{t}+n},a';\theta^-), & s_{\bar{t}+n}\text{不为目标状态}\end{cases}$$

$\bar{t}$ 为情节开始的时刻,R 是每个时刻的立即奖赏,θ^-为目标网络的参数,θ为行为网络的参数.而在情节式 Q 学习中,为了达到提高网络训练过程稳定性的目的,网络参数在一个情节结束后更新,因此仅需使用一个行为网络模型更新网络参数即可.在情节式 Q 学习算法中,当智能体到达目标状态或总时间步数达到最大步长限制时,一个情节终止,即每个情节的持续时间是动态变化的.计算每一时间步的损失函数和梯度,累积梯度,以此更新可训练的网络参数.情节中每一时间步的损失函数为 $(G_{\bar{t}+i}-Q(s_{\bar{t}+i},a;\theta))^2$, 其中, $G_{\bar{t}+i}\leftarrow R_{\bar{t}+i+1}+\gamma G_{\bar{t}+i+1}$,$i\in\{0,1,2,\ldots,T-\bar{t}-1\}$.累积奖赏 G 的初始值定义为:$G_T\leftarrow 0$,T 为情节结束的时刻.

1.3.2 加权双 Q 学习

在 Q 学习[20]的计算过程中,算法使用单估计器来估计状态-动作值,即使用最大状态-动作值来估计最大期望状态-动作值的近似值,导致算法在随机环境中出现值被过高估计的现象.双 Q 学习[21]采用双估计器来避免出现值被过高估计的现象,该算法在确定最优动作及在估计这个动作的状态-动作值时使用了两个经验集(样本集合)互相独立的估计器,会经常出现值被过低估计的现象.加权双 Q 学习是一种基于加权双估计器的算法,其目的是要在过高估计和过低估计之间达到平衡.加权双 Q 学习使用了两个状态-动作值函数(Q^U 和 Q^V)进行计算.对于每一时间步的动作,算法基于这两个状态-动作值函数的线性组合,采用ε贪心策略进行选择.若其中一个值函数要进行更新,那么在更新过程中,该值函数中的状态-动作值定义为

$$Q^{U,WDE}(s,a^*)=\beta^U Q^U(s,a^*)+(1-\beta^U)Q^V(s,a^*) \tag{4}$$

其中,$Q^{U,WDE}$ 为采用加权双估计器计算得到的状态-动作值,a^*为根据 Q^U 所得的最优动作.$\beta^U\in[0,1]$为加权函数,其具体定义为: $\beta^U=\frac{|Q^V(s,a^*)-Q^V(s,a_L)|}{c+|Q^V(s,a^*)-Q^V(s,a_L)|}$,$c\geqslant 0$,$a_L$ 为根据 Q^U 所得的最差动作.当$\beta^U=1$,即 $c=0$ 时,$Q^{U,WDE}$ 等同于采用单估计器得到的状态-动作值;当$\beta^U=0$,即 $c\to\infty$时,$Q^{U,WDE}$ 则等同于采用双估计器得到的状态-动作值.

2 主要成果

本节对本文所提出的主要研究成果分别进行了介绍.第 2.1 节中介绍了广义异步值迭代网络中所用的基于状态的异步更新方法的两种实现形式及主要思想,并对网络中的一次异步值迭代过程进行了描述.第 2.2 节中介绍了情节式加权双 Q 学习算法的主要思想.第 2.3 节介绍了新型图形卷积算子的主要思想.

2.1 广义异步值迭代网络(GAVIN)

基于 GVIN,本文提出了一种异步规划网络模型——GAVIN.该网络利用基于状态的异步更新方法,进一步地改进了 GVIN 中的规划模块,提升了其在具有非规则图形结构任务中的规划性能及其策略在未知任务中的泛化能力.对于输入 GAVIN 的每张非规则图形,网络所采用的基于状态的异步更新方法为图形中每个节点(即 MDP 中的状态)的优先级定义了两种具体形式.在规划模块的每一轮迭代过程中,该方法能根据优先级合理分配各节点的规划时间.

第 1 种形式直接使用贝尔曼误差(Bellman error)来定义节点的优先级.对于 MDP 中的任一状态,其当前贝尔

曼误差为该状态在这轮值迭代前后状态值之差的绝对值,即在一个 MDP 模型已知的环境中,经过第 n 轮值迭代之后,状态 s 的贝尔曼误差 $BE_n(s)$为

$$BE_n(s)=|V_n(s)-\max_a[R(s,a)+\gamma\sum_{s'}Tr(s'|s,a)V_n(s')]|=|V_n(s)-\max_a Q_n(s,a)|=|V_n^{be}(s)-V_n^{af}(s)| \tag{5}$$

其中,$V_n^{be}(s)$表示状态 s 在第 n 轮值迭代之前的状态值,V_n^{af} 表示状态 s 经过了第 n 轮值迭代之后的状态值.

因此,对于 GAVIN 中节点优先级的第 1 种定义方式,在第 n 轮异步值迭代中,当前节点 s 的优先级 $I_n(s)$为

$$I_n(s)=BE_n(s) \tag{6}$$

上述优先级的定义形式基于如下观察:在两轮值迭代之间,有一些节点的状态值会发生显著的变化,因此与这些节点相连接的节点的状态值同样也可能会发生较大的变化.这就意味着:随着值迭代过程的进行,节点的状态值的显著变化会给整个状态空间上与其相连通的节点的状态值带来不同程度的影响.根据贝尔曼误差的定义,节点的状态值的变化越大,贝尔曼误差也就越大,即表明节点的贝尔曼误差可被用于定义优先级——对于那些有着更大贝尔曼误差的节点,在值更新过程中,应赋予它们更高的优先级来优先更新它们的状态值.

第 2 种定义与第 1 种定义略微不同,第 2 种形式中使用转移概率和贝尔曼误差的乘积来定义优先级.对于这种定义方式,在第 n 轮异步值迭代过程中,当前节点 s 的优先级 $I_n(s)$为

$$I_n(s)=TBE_n(s)=Tr(s|s',a')\cdot BE_n(s') \tag{7}$$

其中,$TBE_n(s)$表示第 n 轮异步值迭代过程中,节点 s 上转移概率与贝尔曼误差的乘积.s'是当前节点 s 的前继节点(predecessor node),即能与当前节点之间发生状态转移的节点.s 是节点 s'经过第 n 轮异步值迭代之后能转移到的节点,$Tr(s|s',a')$是智能体在节点 s'执行动作 a'转移到节点 s 的概率.由于在公式(7)的定义中考虑的是两个节点之间转移概率的数值大小而非图形中节点的组成结构,因此利用 $Tr(s|s',a')$而非 $\boldsymbol{P}_{s',s}$ 来表示节点 s'到节点 s 的转移概率.由于非规则图形的结构特性,相互之间能发生状态转移的节点必是相连的节点,所以只要 $Tr(s|s',a')\neq 0$,s'必为 s 的前继节点.

第 1 种优先级定义方式中并没有考虑当前节点与其前继节点之间的转移模型,而在第 2 种定义方式中,为了能更为突出节点之间的连接性,我们引入了“转移”的概念.第 2 种定义方式的主要思想与第 1 种定义方式的思想类似,具体为:若状态空间中某些节点(如节点 s')的状态值在值迭代前后的变化越大,那么那些能与其发生状态转移的节点(如节点 s)的状态值发生的变化也会越大.这就意味着:随着迭代过程的执行,节点 s 的状态值的变化会对与其之间有着较大转移概率 $Tr(s|s',a')$的前继节点 s'的状态值带来较大的影响.因此,在值更新过程中,应该赋予这些节点 s 较高的优先级来优先更新它们的状态值.在 GAVIN 中,无论是利用第 1 种方式还是第 2 种方式来定义节点的优先级,只要节点的状态值随着迭代的进行发生了变化,那么该节点的优先级也会随之改变.

在定义了节点的优先级之后,就可根据优先级来选择每轮异步值迭代中要进行值更新的节点.为了能合理地选择节点,本文根据各节点的优先级定义了一个阈值,并在每轮迭代开始前选择那些优先级大于阈值的节点进行更新.本文使用所有节点贝尔曼误差的平均值作为阈值,也就是说,第 n 轮异步值迭代过程开始前,阈值 T_n 的定义为

$$T_n=\frac{1}{|\nu|}\sum_{s\in\nu}I_n(s) \tag{8}$$

其中,ν表示一张非规则图形的整个节点空间,s 表示图中的任一节点,$|\nu|$表示图中的节点总数.

由公式(8)可知,该阈值在不同轮次的异步值迭代中的大小也会不同.这就使得在每轮异步值迭代中,所选节点的个数会根据当前环境自适应地变化.此外,使用所有节点贝尔曼误差的平均值作为阈值,能够使得那些具有相对较高优先级(如优先级高于 T_n)的节点与那些具有相对较低优先级(如优先级低于 T_n)的节点更具区分性.

根据节点的优先级和阈值,选择第 n 轮异步值迭代过程中要进行值更新的节点的过程可形式化为

$$S_n^E=f_\nu(\nu;T_n) \tag{9}$$

其中,f_ν为节点选择函数,S_n^E 表示被选择出来以执行值更新的节点集合.在选择好要进行值更新的节点后,即可执行 GAVIN 中的异步值迭代过程.GAVIN 中的第 n 轮异步值迭代的过程可被形式化为

$$\boldsymbol{Q}_{n+1}^{(a)}(S_n^E)=\boldsymbol{P}^{(a)}(S_n^E)(\boldsymbol{R}+\gamma\boldsymbol{V}_n) \tag{10}$$

$$\boldsymbol{V}_{n+1}(S_n^E)=\max_a \boldsymbol{Q}_{n+1}^{(a)}(S_n^E) \tag{11}$$

其中, $\boldsymbol{P}^{(a)}(S_n^E)$ 为第 a 个通道上的用于更新所选节点值函数的图形卷积算子.

图 2 表示的是 GAVIN 中一次异步值迭代的过程.为了简化该过程示意图,本文对图中输入网络的非规则图形的结构进行了简化.图中左下角的非规则图形 G 为输入到所示网络的 5-节点的非规则图形.图中乘法符号代表的操作为两个矩阵相乘,加法符号代表的操作为两个矩阵相加.该过程的具体解释如下:在异步值迭代过程开始前,先利用函数 f_v选择出那些优先级大于阈值的节点,并将其表示为一组深灰色方块,即 S_n^E. 在图形卷积算子的每个通道中,那些能与所选节点发生状态转移的操作也相应地被选择出来,并在图中以带阴影的方块标记.根据所选节点来选择图形卷积算子的操作,在图中以带箭头的虚线表示.利用被选择出来的图形卷积算子 $\boldsymbol{P}(S_n^E)$, 图形的值信号 $\boldsymbol{R}+\gamma\boldsymbol{V}_n$ 得以在图中有目的性地扩散,并最终获得所选节点的状态-动作值图形信号 $\boldsymbol{Q}_{n+1}(S_n^E)$.这个操作相当于图卷积的过程,以矩阵相乘的方式执行.通过对 $\boldsymbol{Q}_{n+1}(S_n^E)$ 各个通道上的值取最大值,获得所选节点更新后的状态值 $\boldsymbol{V}_{n+1}(S_n^E)$. 该操作相当于最大池化的过程.优先级小于阈值的节点在图中被表示为浅灰色方块,且它们的状态值 $\boldsymbol{V}_n(\complement_v S_n^E)$ 不会被更新($\complement_v S_n^E$ 表示 S_n^E 的补集).在一轮异步值迭代之后,整个节点空间的状态值 $\boldsymbol{V}_{n+1}$ 由更新后的节点状态值 $\boldsymbol{V}_{n+1}(S_n^E)$ 及未更新的节点状态值 $\boldsymbol{V}_n(\complement_v S_n^E)$ 组成.

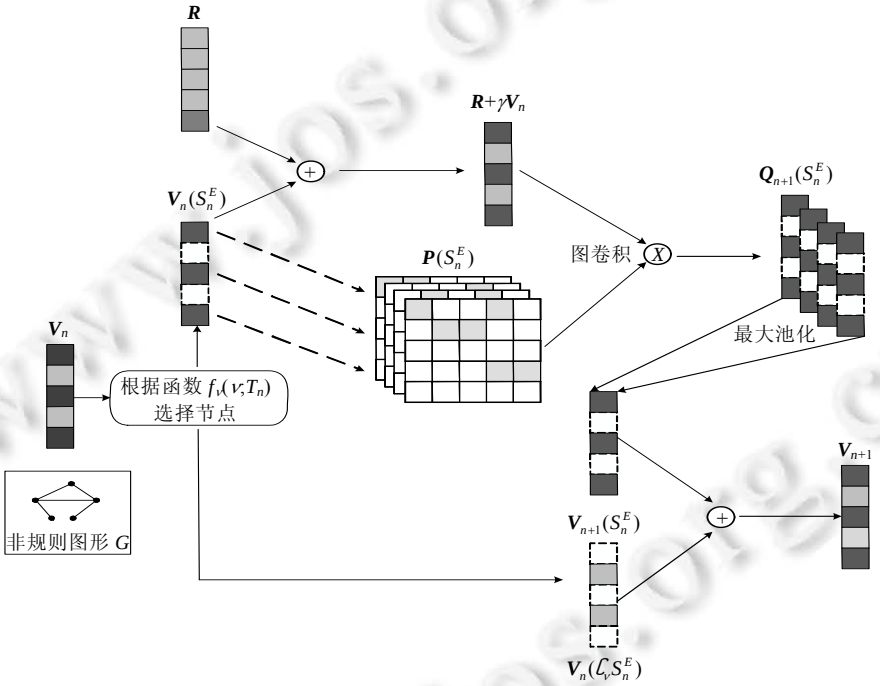

Fig.2 An asynchronous value iteration process in GAVIN

图 2 GAVIN 中的一次异步值迭代过程

2.2 情节式加权双Q学习

与传统 Q 学习算法相比,尽管情节式 Q 学习能够提升规划网络模型的训练稳定性,但在网络参数的更新过程中,其仍会出现 Q 学习算法中常见的值被过高估计的问题[22].与 Q 学习算法相同,情节式 Q 学习在选择最优动作和计算目标值时使用了相同参数的模型,因此在计算目标值时会给出一个状态-动作值的上限的估计值.由于模型在训练过程中并不够稳定,该估计值可能存在一些偏差.如果这个偏差不一致,则会造成模型对动作优劣的判断产生失误,从而影响训练出的模型的性能.为了在保持网络模型的训练稳定性的同时减轻过高估计对训练性能的影响,本文结合加权双估计器及情节式 Q 学习,提出了情节式加权双 Q 学习算法.

情节式加权双 Q 学习算法将被作用于规划网络模型中,其伪代码如算法 1 所示.解释如下:当一个情节开始时,对于每个给定的起始节点 s_0,智能体会同时考虑行为网络 $\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w}')}$ 及目标网络 $\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w})}$,即 $\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w}')}+\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w})}$,并基于这个线性组合,采用$\varepsilon$贪心策略选择动作,使其从当前节点 s_t 到达下一个节点 s_{t+1}.也就是说:在智能体选择下一个节点过程中,存在着$(1-\varepsilon)$的概率,使得 $s_{t+1}=\arg\max_{s'\in Nei(s_t)}(\boldsymbol{Q}_{s'}^{(\boldsymbol{w})}+\boldsymbol{Q}_{s'}^{(\boldsymbol{w}')})$;存在着$\varepsilon$的概率,$s_{t+1}$ 是从与 s_t 相连的节点中随机选择的.算法的输入为非规则图形,图形中每个节点都具有不同的局部结构,因而智能体在每个节点上可执行的动作数目也是不同的.在训练期间,需使用一个伪状态-动作值来表示智能体从节点 s 转移到与其相连的某个节点时所对应的状态-动作值,具体定义为 $\boldsymbol{Q}_s=\max_{s'\in Nei(s)}\boldsymbol{V}_{s'}$,其中,$Nei(s)$表示与节点 s 相连的节点集合,$\boldsymbol{V}_{s'}$表示与节点 s 相连的节点的状态值.当 s_t 是目标节点或总步数达到最大步长限制时,一个情节结束.当一个情节结束时,构建函数 $\beta=\dfrac{|\boldsymbol{Q}_{s_t}^{(a^*;\boldsymbol{w})}-\boldsymbol{Q}_{s_t}^{(a_L;\boldsymbol{w})}|}{c+|\boldsymbol{Q}_{s_t}^{(a^*;\boldsymbol{w})}-\boldsymbol{Q}_{s_t}^{(a_L;\boldsymbol{w})}|}$ 以对目标值进行加权处理,其中,$c\geqslant 0$,a^*是基于行为网络 $\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w}')}$ 的最优动作,a_L 是基于行为网络 $\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w}')}$ 的最差动作.此操作旨在让模型在选择最大动作和计算目标值时使用不同的网络参数.若情节终止时,智能体所在的节点 s_t 为目标节点,则初始化回报,即目标值,$G=0$;否则,G 的初始值为 $\boldsymbol{Q}_{s_t}^{(a^*;\boldsymbol{w}')}$ 与 $\boldsymbol{Q}_{s_t}^{(a^*;\boldsymbol{w})}$ 的加权线性组合.最后考虑损失函数 $\boldsymbol{L}(\boldsymbol{w}')=\sum_{t=1}^{T}(G_t-\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w}')})^2$,$G_t$ 为智能体在 t 时刻的回报,定义为 $G_t=(R_{t+1}+\gamma G_{t+1})$,其中,$R_t$ 是智能体在 t 时刻所获的立即奖赏.计算完每一时间步的损失函数和梯度后,累积整个情节的梯度,并利用其来更新目标网络的参数.

算法 1. 情节式加权双 Q 学习.

输入:一张带有目标节点 s_g 的非规则图形 G.

1. 初始化:情节数 $T=0$,目标网络参数 $\boldsymbol{w}$,行为网络参数 $\boldsymbol{w}'$,网络梯度$\Delta\boldsymbol{w}$,情节内步数 $t=0$
2. **REPEAT**:
3. 清除网络梯度$\Delta\boldsymbol{w}\leftarrow 0$;
4. 行为网络参数 $\boldsymbol{w}'=\boldsymbol{w}$;
5. 随机选择一个起始节点 s_t;
6. **REPEAT**:
7. 基于 $\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w}')}+\boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w})}$,根据$\varepsilon$贪心策略选择动作;
8. 获取奖赏 R_t 并到达下一个节点 s_{t+1};
9. $t=t+1$;
10. **UNTIL** 到达终止条件 $s_t=s_g$ 或 $t>t_{\max}$;
11. $a^*\leftarrow\arg\max_a \boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w}')}$;
12. $a_L\leftarrow\arg\min_a \boldsymbol{Q}_{s_t}^{(a;\boldsymbol{w}')}$;
13. $\beta\leftarrow\dfrac{|\boldsymbol{Q}_{s_t}^{(a^*;\boldsymbol{w})}-\boldsymbol{Q}_{s_t}^{(a_L;\boldsymbol{w})}|}{c+|\boldsymbol{Q}_{s_t}^{(a^*;\boldsymbol{w})}-\boldsymbol{Q}_{s_t}^{(a_L;\boldsymbol{w})}|}$;
14. $G\leftarrow\begin{cases}0, & \text{到达目标节点}s_g\\ \beta\boldsymbol{Q}_{s_t}^{(a^*;\boldsymbol{w}')}+(1-\beta)\boldsymbol{Q}_{s_t}^{(a^*;\boldsymbol{w})}, & \text{未到达目标节点}s_g\end{cases}$
15. **FOR** $i=t:-1:0$:
16. $G\leftarrow R_i+\gamma G$;
17. 累积梯度 $\Delta\boldsymbol{w}\leftarrow\Delta\boldsymbol{w}+\dfrac{\partial(G-\boldsymbol{Q}_s^{(a;\boldsymbol{w}')})^2}{\partial\boldsymbol{w}'}$;
18. **END FOR**
19. $\boldsymbol{w}\leftarrow\boldsymbol{w}-\Delta\boldsymbol{w}$;

20. $T=T+1$;

21. **UNTIL** $T>T_{\max}$

在算法 1 中可看出,情节式加权双 Q 学习仍保留了在情节结束时更新网络权重的属性.但与原算法不同,情节式加权双 Q 学习算法引入了行为网络和目标网络来共同完成网络参数的更新.算法利用行为网络 $\boldsymbol{Q}^{w'}$来选择最优动作和最差动作,并根据这两个动作,利用目标网络 $\boldsymbol{Q}^{w}$构建加权函数β.通过这个加权函数,对行为网络 $\boldsymbol{Q}^{w'}$和目标网络 $\boldsymbol{Q}^{w}$进行加权线性求和,并将这个加权线性组合作为目标值来计算损失函数,以获取梯度来更新可训练的网络权重.这一操作不仅有效缓解了原算法中存在的值过高估计问题,同时还避免了直接使用双 Q 学习算法时可能面临的值被过低估计问题,最终在值被过高估计与值被过低估计之间达到平衡[18].因此,在使用情节式加权双 Q 学习算法训练规划网络时,智能体可使用一个更有效且更稳定的规划方案来完成每个训练情节.

2.3 新型图形卷积算子

在 GVIN 及 GAVIN 中,图形卷积算子 $\boldsymbol{P}$ 由邻接矩阵 $\boldsymbol{A}$、图形中各节点嵌入信息 $\boldsymbol{X}$ 以及核函数共同确定.对于任意一张图,邻接矩阵即反映了图中各节点的度(与节点相关的边的数目)的分布情况.这就意味着,该图形卷积算子将受到图中每个节点的度的分布的影响.在各节点进行信息交互及信息传递的过程中,节点的度决定了该节点在该过程中的重要性,即决定了该节点所能接收的信息量.因此,网络中基于图卷积操作的规划结果也会受到节点的度的分布的影响[23−25].而在规划过程中,由于动作的目标导向性,目标节点及其附近的节点应是更为重要的,但如果输入网络的非规则图形中远离目标节点的某些节点具有相对较大的度的话,那么由于图形卷积算子的影响,在规划过程中,图形中其他节点上的信息转移到这些具有较大度的节点的概率就会越大,这可能会使得智能体在规划过程中混淆各节点的重要性,从而导致具有较大度的节点会具有相对较大的状态值,有时会超过与目标节点相连的节点甚至是目标节点的状态值,导致网络最终规划性能的下降.

为了尽可能避免网络的规划过程中出现上述这一现象,本文对 GVIN 中提出的由基于嵌入信息的核函数定义的图形卷积算子进行了改进.由于核函数 $K_{emb}(\boldsymbol{X}_i,\boldsymbol{X}_j)$关注的是如何准确学习到图中任意两个节点之间的隐藏结构信息,其并没有考虑到整个非规则图形中节点的组成结构,因此本文并不考虑对核函数进行改动.从公式(3)的定义可看出,在这种类型的图形卷积算子 $\boldsymbol{P}$ 中,图形的邻接矩阵利用一个分数形式的操作进行了归一化,并形成一个激活系数,该激活系数利用图形中各节点的连接性来激活核函数中相连节点之间的转移信息.但这个分数形式的激活系数定义存在着一个弊端——它会使得任意两个相连节点之间的转移概率相同,即 $\boldsymbol{P}_{i,j}=\boldsymbol{P}_{j,i}$.如果两个相连的节点中的其中一个节点的度较大,而另一个节点的度较小时,利用这种形式的激活系数可能会造成不公平的转移概率分配,从而出现上文所述的严重后果.本文基于这个激活系数对原有的图形卷积算子进行了改进,在新的图形卷积算子中,(i,j)节点之间转移概率为

$$\boldsymbol{P}'_{i,j}=\frac{I_{i=j}+\boldsymbol{A}_{i,j}}{\left[\sum_k(I_{i=k}+\boldsymbol{A}_{i,k})\right]^{\alpha}}\cdot K_{emb}(\boldsymbol{X}_i,\boldsymbol{X}_j) \tag{12}$$

其中,α=1.与原始的图形卷积算子不同,改进后的图形卷积算子 $\boldsymbol{P}'$仅考虑了单个节点的节点度.对于任意两个节点,利用这个新定义的激活系数不仅能确定节点之间的连接性,还能够根据它们各自的度来合理分配节点之间转移概率的大小,即对于度相对较大的节点,图中其他节点上的信息转移到它的概率就相对较小.网络利用这一改进后的图形卷积算子进行规划运算可有效地弱化图形中节点的度的分布对规划结果的影响,进而提高网络的规划性能以及其在未知任务中的泛化能力.

3 实验结果及分析

在不同规模的非规则图形以及真实路况地图中,本节对所提出的广义异步值迭代网络 GAVIN 的训练性能(即在训练集中的性能)、泛化能力(即在测试集中的性能)及规划性能(即在真实路况地图中的性能)进行了全面评估,并对情节式加权双 Q 学习及新型图形卷积算子的有效性进行了验证.

3.1 实验环境及参数设置

给定起始节点和目标节点,本节中的实验考虑如何使得智能体在具有非规则图形结构的环境中规划出一条或多条能够成功到达目标节点的最优路径.要注意的是:最优路径可以是自定义的,不一定为最短路径.用于进行实验的数据集分别为非规则图形(10-节点、100-节点)以及真实路况地图(明尼苏达高速地图、纽约市区街道地图,实验中所使用的非规则图形及真实路况地图数据集的来源为:https://github.com/sufengniu/GVIN/tree/master/data).图 3 表示了实验中所用的 3 种数据集.网络中的规划模块被表示为特定类型的卷积神经网络,网络中编码的网络参数可以通过反向传播算法进行训练.所有实验均使用学习率为 η=0.001 的标准 RMSProp 算法作为优化器,RMSProp 衰减因子为 0.999.对于利用 IL 算法作为训练算法的网络,其具体训练方法为采用特定数据集中的专家样本进行训练;对于利用 RL 算法训练的网络,训练算法为情节式 Q 学习算法和情节式加权双 Q 学习算法.所有实验中所用的图形卷积算子均使用基于节点嵌入信息的核函数进行定义,该核函数的结构是一个 3 层全连接神经网络(32-64-1),每层均使用 $ReLU(\cdot)=\max(0,\cdot)$作为激活函数,网络权重使用期望为 0、方差为 0.01 的正态分布进行初始化.所有网络所用的图形卷积算子的通道数目均被设置为 10.在网络的测试过程中,所有实验的立即奖赏设置相同,即在每个时间步之后,除了已到达目标节点的情况外,智能体会获得一个与其步长相关联的负奖赏$-0.1*L$,其中,L 表示每一时间步的步长.当在限定步数之内到达目标节点时,智能体会获得一个+1 的奖赏;而在网络的训练过程中,实验中增大了智能体在每一时间步所获的负奖赏,将其变为$-5\times L$,而智能体到达目标节点时所获的正奖赏仍保持不变.为了增大训练及测试的难度以突显 GAVIN 以及利用情节式加权双 Q 学习算法进行训练的网络的优势,本文实验中的一个重要参数的设置与文献[12]中的设置不同,即奖赏值:在文献[12]中,无论是训练还是测试过程,智能体在每一时间步的奖赏值均为$-0.01\times L$,而到达目标节点的奖赏值仍为+1.因此,本文所获得的实验结果也会与文献[12]中的数据有一定的差异.

(a) 100-节点的非规则图形 (b) 可视化后的明尼苏达高速地图 (c) 可视化后的纽约市区街道地图

Fig.3 Examples of three types of data sets used in the visualization experiments

图 3 可视化实验中所用的 3 种数据集示例

网络的性能使用 3 个指标来进行量化,分别为成功率、期望回报和更新次数.成功率是指:在其所采取的当前步数超过最大步数限制之前,智能体能够成功地从起始节点到达目标节点的概率.这个指标反映了网络在训练任务中的规划性能以及在未知的测试任务中的泛化能力.期望回报,即策略的期望累积奖赏,其大小与网络所规划出的路径长度直接相关.无论是在训练任务还是在测试任务中,该指标均反映了网络所规划的策略的质量.成功率和期望回报越高,网络的规划性能及泛化能力就越好.更新次数是指网络在训练时,完成整个规划过程所需进行值更新的节点数目,该指标可用于比较网络的规划效率.

3.2 情节式Q学习算法下的网络训练性能对比

该实验利用情节式 Q 学习算法分别训练 GVIN 和 GAVIN,并评估了它们的训练结果.实验中,用于进行训练的数据集的类型为 10-节点的非规则图形.在网络规划模块中,值迭代过程的循环次数 N=15.训练集大小分别为 1 428 张、4 285 张、8 571 张,各占数据集大小的 1/7、3/7、6/7.每个训练集中均设置了 5 个不同的种子用于初始化网络模型,最终训练结果为这 5 个模型训练结果的平均值.

图 4 展示了在不同大小训练集下,利用不同的优先级方法得到的广义异步值迭代网络(GAVIN-BE/GAVIN-TBE)以及 GVIN 在情节式 Q 学习算法下的训练结果.GAVIN-BE 为采用第 1 种优先级方法的广义异步值迭代网络;GAVIN-TBE 为采用第 2 种优先级方法的广义异步值迭代网络.

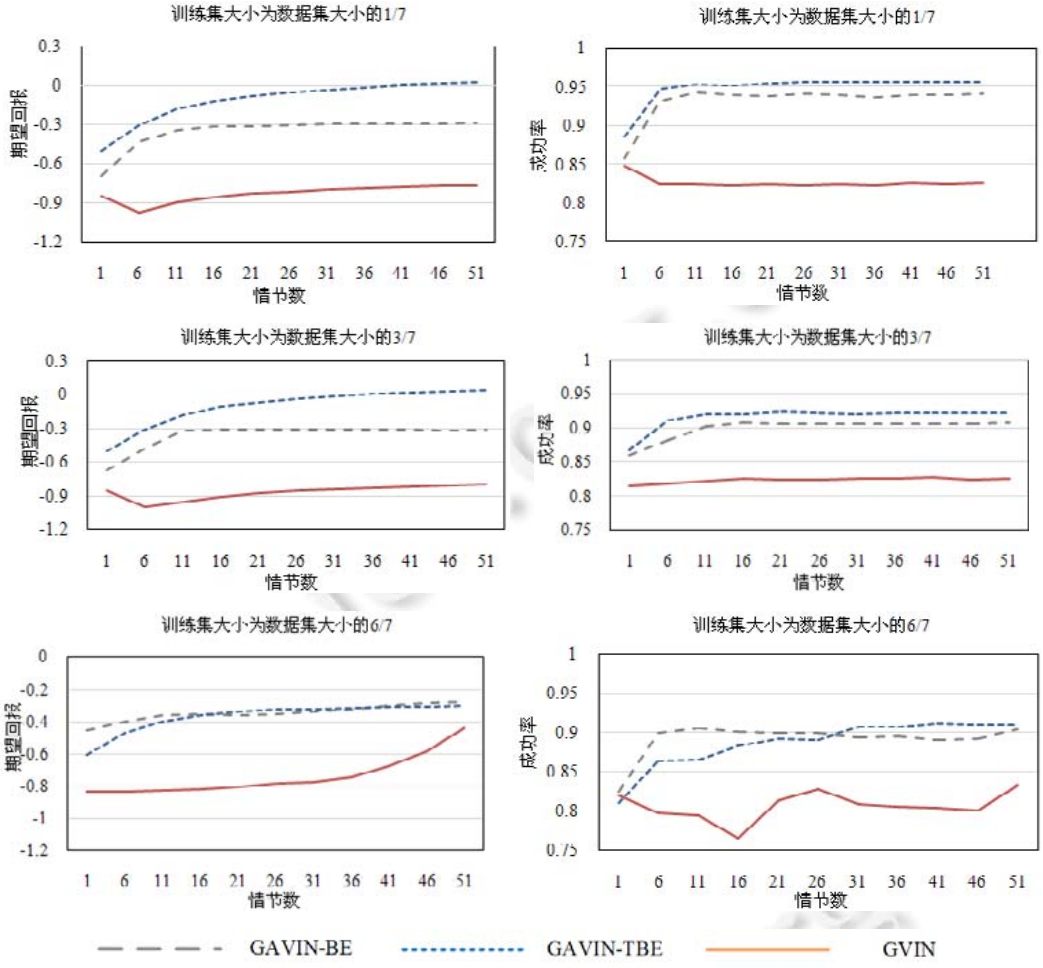

Fig.4 Comparison of training performance between GVIN and GAVIN (GAVIN-BE/GAVIN-TBE) under training sets with different sizes

图 4 在不同大小的训练集中,GVIN 与 GAVIN(GAVIN-BE/GAVIN-TBE)的训练性能对比

如图 4 所示,从成功率及期望回报这两个指标来看,无论是利用 1 428 张、4 285 张还是 8 571 张 10-节点的非规则图形进行训练,GAVIN 的训练结果均远优于 GVIN.这是由于在值更新的过程中,GAVIN 会更关注于对目标节点以及那些与目标节点密切相关连的节点的状态值进行更新.这一特性增强了 GAVIN 规划过程中智能体所执行的动作的目标导向性,从而使其能比 GVIN 训练得更好.而由于训练难度较大,利用这一奖赏值设定进行训练的 GVIN 的训练性能很难随着学习过程而提升,且当训练数据量较少(训练集为 1 428 张)时,还会出现性能下降和无法收敛的问题.这一结果表明:即使在训练条件较为恶劣的实验环境中,GAVIN 仍能比 GVIN 获得更好的训练性能.对于采用不同优先级定义方法的 GAVIN,GAVIN-TBE 的训练性能均会略优于 GAVIN-BE.这一现象表明:即便这两种优先级方法均能使得 GAVIN 比 GVIN 获得更好的规划性能及训练性能,但相对而言,在节点优先级的定义中考虑节点之间的转移模型,能使得网络中所执行的异步值更新过程更为有效.

3.3 不同RL训练算法下的网络训练性能对比

该实验分别使用情节式 Q 学习算法及情节式加权双 Q 学习算法训练 GVIN,并比较不同训练算法下 GVIN 的训练性能.实验中,用于训练的数据集为 10-节点的非规则图形,训练集有 8 571 张,占数据集大小的 6/7;且在网络规划模块中,设置值迭代过程的循环次数 N=15.每个训练集中均设置了 5 个不同的种子用于初始化网络模型,最终训练结果为这 5 个模型训练结果的平均值.

图 5 中的实验结果展示了利用情节式 Q 学习和情节式加权双 Q 学习作为训练算法的 GVIN 的训练性能.在情节式加权双 Q 学习中,为了验证加权函数β的大小对于算法性能的影响,本文为参数 c 设置了 3 个不同大小的值,分别为 1,10,100.根据算法 1 中函数β的定义可知,c 越大,算法就越接近于使用双估计器的算法;c 越小,算法就越接近于使用单估计器的算法.如图 5 所示,从成功率及期望回报这两个指标来看,利用情节式加权双 Q 学习算法训练的网络会远好于利用情节式 Q 学习算法训练的网络.这一结果说明:利用情节式加权双 Q 学习算法进行训练能使得网络的规划过程更为有效,从而获得更好的训练性能.除此之外,从利用情节式加权双 Q 学习算法所得到的网络的训练结果中,随着训练过程中情节数的增多,当 c=1 时,由于近似于使用单估计器,网络会产生过高估计的现象;在 c=100 时,由于近似于使用双估计器,网络会产生过低估计的现象;在 c=10 时,网络相当于使用加权双估计器进行计算,能较好避免过高和过低估计的问题,因而获得更好的训练效果.后续实验中,将使用 c=10 得到的网络模型进行测试.

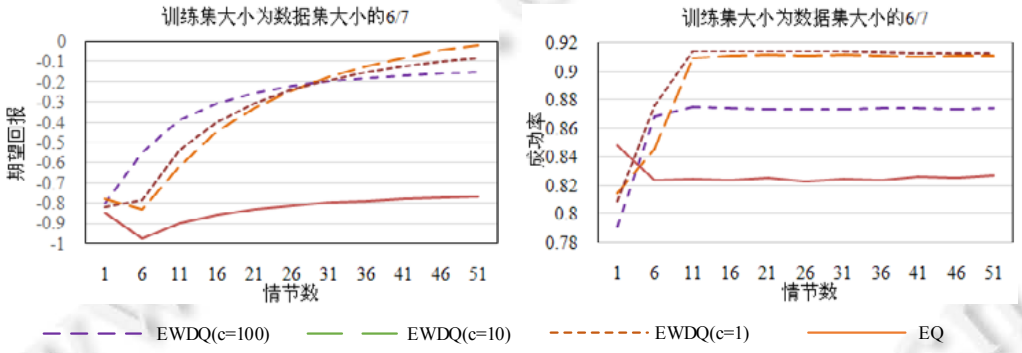

Fig.5 Comparison of training performance between GVIN using episodic Q-learning (EQ) and GVIN using episodic weighted double Q-learning (EWDQ)

图 5 利用情节式 Q 学习(EQ)和情节式加权双 Q 学习(EWDQ)得到的 GVIN 的训练性能对比

3.4 不同算法在非规则图形中的网络泛化能力对比

该实验分别利用经过 IL 算法及 RL 算法训练后的 GVIN 及 GAVIN 模型,在与训练集不同且结构更为复杂的非规则图形中进行测试,并通过测试结果比较网络的泛化能力.实验中:用于进行训练的数据集为 10-节点的非规则图形,训练集为 8 571 张,占数据集大小的 6/7;用于进行测试的数据集为 100-节点的非规则图像,测试集为 1 428 张,占数据集大小的 1/7.对于利用 IL 算法作为训练算法的网络,在网络规划模块中,设置值迭代过程的循环次数 N=30;对于利用 RL 算法训练的网络,在网络规划模块中,设置值迭代过程的循环次数 N=15.该循环次数的设置,能使网络充分更新节点值函数并训练出较好的模型.每个训练集中均设置了 5 个不同的种子用于初始化网络模型,并在训练完成后,利用这 5 个网络模型分别进行测试.最终测试结果为这 5 个网络模型测试结果的平均值.

根据表 1 及表 2 中的数据对比可得:无论是利用 IL 算法还是 RL 算法进行训练,在未知的、且结构更复杂的非规则图形上进行测试时,GAVIN 的成功率和期望回报都远优于 GVIN;同时,前者进行训练时所需的节点更新次数还远少于后者.这一结果说明:尽管在每一轮异步值迭代过程中,GAVIN 中需更新的节点个数远少于 GVIN 中需更新的节点个数,但前者所更新的节点均为那些能给网络的规划性能带来显著提升的节点,这使得

GAVIN 中所执行的规划过程更高效且更有效,从而具有更好的泛化能力.同时,从表 2 中的第 1 列及第 3 列数据对比可看出,采用情节式加权双 Q 学习算法进行训练的网络,不仅能在训练性能上远优于采用情节式 Q 学习算法进行训练的网络,而且在网络的泛化能力方面也具有同样的优势.这一结果有力地验证了本文中所提出的情节式加权双 Q 学习算法的有效性.除此之外,从表 1 中第 3 列、第 4 列及表 2 中第 4 列、第 5 列的数据对比可看出,对于不同的优先级方法,GAVIN-TBE 的泛化能力会略优于 GAVIN-BE.结合第 3.2 节中的训练结果来看,在优先级定义中考虑节点之间的转移模型,能够使得网络中所执行的异步值更新过程更为有效.

Table 1 Comparison of generalization abilities between GVIN using different graph convolution operators and GAVIN (GAVIN-BE/GAVIN-TBE) under IL training algorithms

表 1 IL 训练算法下,使用不同图形卷积算子的 GVIN 及 GAVIN(GAVIN-BE/GAVIN-TBE)的泛化能力对比

性能指标	网络模型			
	GVIN	GVIN-newP	GAVIN-BE	GAVIN-TBE
成功率(%)	60.70	**64.81**	**78.38**	**81.72**
期望回报	0.552 6	**0.593 3**	**0.746 4**	**0.784 7**
更新次数	300	300	**138**	**124**

Table 2 Comparison of generalization abilities between GVIN using different graph convolution operators and GAVIN (GAVIN-BE/GAVIN-TBE) under RL training algorithms

表 2 RL 训练算法下,使用不同图形卷积算子的 GVIN 及 GAVIN(GAVIN-BE/GAVIN-TBE)的泛化能力对比

性能指标	网络模型				
	GVIN	GVIN-newP	GVIN-EWDQ	GAVIN-BE	GAVIN-TBE
成功率(%)	62.97	**77.71**	**77.82**	**82.43**	**85.32**
期望回报	0.577 8	**0.740 5**	**0.752 9**	**0.806 1**	**0.847 7**
更新次数	150	150	150	**23**	**30**

3.5 不同图形卷积算子下的网络泛化能力对比

该实验在GVIN上对改进后的图形卷积算子的性能进行评估.除了图形卷积算子中激活系数的定义与原始图形卷积算子中的定义不同之外,该实验中的其他实验设置均与第 3.4 节中的实验设置相同.

表 1 中的第 2 列及表 2 中的第 2 列数据分别表示的是采用了改进后的图形卷积算子的 GVIN 在 IL 及 RL 算法训练后的测试结果.经由这两列数据与表 1 中的第 1 列及表 2 中的第 1 列数据对比可得:采用了改进后的图形卷积算子的网络在未知的测试任务上的泛化能力明显优于采用原始图形卷积算子的网络,且相比于利用 IL 训练算法进行训练的网络,在 RL 训练算法下,改进后的图形卷积算子为网络泛化能力带来的提高更为明显.这一结果说明:本文中所提出的改进后的图形卷积算子能够有效地解决原图形卷积算子中存在的转移概率分配不公平的问题,进而提高了网络的规划能力,使得网络产生的策略可以在未知的测试任务中获得更好的泛化能力.在利用 RL 训练算法所得到的网络中,由于转移概率对算法性能的影响较大,因此该改进后的图形卷积算子给网络性能带来的提高也会尤为明显.

3.6 不同算法在真实路况地图中的网络泛化能力对比

该实验利用经过情节式加权双 Q 学习算法训练后的 GVIN 及 GAVIN 模型,在真实路况地图中进行测试,并通过测试结果比较网络在大规模实验应用场景中的泛化能力.其中,GAVIN 中的优先级定义方法为第 2 种方法,即 GAVIN-TBE.实验中:用于进行训练的数据集为 100-节点的非规则图形,训练集为 8 571 张,占数据集大小的 6/7;用于进行测试的数据集分别为明尼苏达高速地图以及纽约市区街道地图.其中,明尼苏达高速地图包含了 2 632 个用以表示路口的节点以及 6 606 条用以表示道路的边,纽约市区街道地图包含了 5 069 个用以表示路口的节点以及 13 368 条用以表示道路的边.实验中 GVIN 和 GAVIN 均采用了改进后的图形卷积算子进行训

练和测试,且在网络规划模块中,设置值迭代过程的循环次数 N=150.每个训练集中均设置了 5 个不同的种子用于初始化网络模型,并在训练完成后,利用这 5 个网络模型分别进行测试.最终测试结果为这 5 个网络模型测试结果的平均值.

根据表 3 中的数据对比可得:在未知的大规模真实路况地图上,利用在 100-节点的非规则图形中训练好的模型进行测试时,GAVIN 的成功率和期望回报都远优于 GVIN;同时,前者进行训练时所需的节点更新次数还远少于后者.这一结果说明:通过利用基于状态的异步更新方法,本文所提出的 GAVIN 在内部组成结构非常复杂的大规模实际应用场景中,能够根据状态的优先级更好地规划出一条或多条成功到达目标点的路径,同时还能够保证较高的规划效率.以上结论有力地验证了 GAVIN 在大规模实际应用场景中的适用性和有效性,这也充分表明,这个改进的规划网络模型能为许多实际应用场景带来益处.

Table 3 Comparison of generalization abilities between GVIN and GAVIN in real road maps under EWDQ training algorithm

表 3 情节式加权双 Q 学习训练算法下,GVIN 以及 GAVIN 在真实路况地图中的泛化能力对比

性能指标	明尼苏达高速地图		纽约市区街道地图	
	GVIN	GAVIN-TBE	GVIN	GAVIN-TBE
成功率(%)	72.31	**84.63**	61.59	**81.73**
期望回报	0.713 8	**0.831 3**	0.593 1	**0.798 2**
更新次数	15 000	4 271	15 000	4 271

图 6 表示了经过 100-节点非规则图形训练后的 GAVIN(左)和 GVIN(右)在明尼苏达高速地图以及纽约市区街道地图中的规划路径对比示例图(六角星表示目标点).从图中结果可明显看出,当给定的目标点与起始点相距较远时,GVIN 所规划出的路径会无法成功达到目标点,而 GAVIN 则能够更好地规划出成功到达目标点的路径.值得注意的是,表 3 中的更新次数在明尼苏达高速地图以及纽约市区街道地图中是相同的.原因为:用于在两个地图环境中进行测试的模型均相同,即都是利用 100-节点的非规则图形训练好的 GVIN 模型或是 GAVIN 模型,而更新次数表示的是模型在训练过程中所需要更新的节点数,所以即便是在不同的两个测试集中进行测试,更新次数均为相同的.

(a) GAVIN 和 GVIN 在明尼苏达高速地图中所规划的路径 (b) GAVIN 和 GVIN 在纽约市区街道地图中所规划的路径

Fig.6 Examples of the planning paths of GAVIN and GVIN in the real road visualization experiments

图 6 GAVIN 和 GVIN 在真实路况可视化实验中的规划路径示例

4 结 论

本文提出了一种可微的广义异步值迭代网络模型——GAVIN.它可利用 IL 或 RL 算法进行端到端的训练,处理具有非规则图形结构的任务,且产生的策略能用于更复杂的未知任务.通过在网络的值更新过程中使用基于状态的异步更新方法,GAVIN 能获得高效且有效的规划过程,从而使其能在未知的非规则结构任务中获得较好的泛化能力.同时,本文还将加权双估计器与情节式 Q 学习算法相结合,提出了一种用于训练网络参数的更高效的 RL 算法——情节式加权双 Q 学习.与原算法相比,该算法显著提升了网络的泛化能力及训练稳定性.此外,

本文提出了一种新型的图形卷积算子,且在实验中验证了它的有效性.

目前,本文所提出的规划网络模型及用于训练网络的 RL 算法仍存在一些不足之处,未来可围绕其做进一步的研究.例如,可寻找一种更好的方法来定义 GAVIN 的异步值迭代过程中各节点的优先级以及用于选择要更新的节点的阈值,使得网络可更好地应用于更大规模且内部组成结构更为复杂的应用场景,从而获得更好的泛化能力.此外,由于 GAVIN 的每轮异步值迭代过程仅会选择特定的节点进行更新,因此在利用 IL 算法进行训练的 GAVIN 的测试结果中会存在一定的过拟合现象,未来可寻求一种更好的神经网络结构来构建模型或是采用数据增强以及数据清洗的方法以消除这一现象.在本文所提出的情节式加权双 Q 学习中,加权函数的大小仍是人为设定的,未来可寻求一种无监督的参数设置方法来自动设定算法中加权函数的大小.

References:

[1] Sun ZJ, Xue L, Xu YM, Wang Z. Overview of deep learning. Application Research of Computers, 2012,29(8):2806−2810 (in Chinese with English abstract).

[2] Liu Q, Zhai JW, Zhang ZZ, Zhong S, Zhou Q, Zhang P, Xu J. A survey of deep reinforcement learning. Chinese Journal of Computers, 2018,41(1):1−27 (in Chinese with English abstract).

[3] Hussein A, Gaber MM, Elyan E, Jayne C. Imitation learning: A survey of learning methods. ACM Computer Survey, 2017,50(2): 21:1−21:35.

[4] Ciresan DC, Meier U, Schmidhuber J. Multi-column deep neural networks for image classification. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2012. 3642−3649.

[5] Farabet C, Couprie C, Najman L, LeCun Y. Learning hierarchical features for scene labeling. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2013,35(8):1915−1929.

[6] Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). 2012. 1106−1114.

[7] Chen XS, Li S, Li H, Jiang SH, Qi Y, Song L. Generative adversarial user model for reinforcement learning based recommendation system. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2019. 1052−1061.

[8] Qureshi AH, Boots B, Yip MC. Adversarial imitation via variational inverse reinforcement learning. In: Proc. of the Int'l Conf. on Learning Representations (ICLR). 2019.

[9] Bertsekas DP. Dynamic Programming and Optimal Control. 3rd ed., Athena Scientific, 2005.

[10] Tamar A, Wu Y, Thomas G, Levine S, Abbeel P. Value iteration networks. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). 2016. 2154−2162.

[11] Bellman R. Dynamic Programming. Princeton: Princeton University Press, 1957.

[12] Niu SF, Chen SH, Guo HY, Targonski C, Smith MC, Kovacevic J. Generalized value iteration networks: Life beyond lattices. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI). 2018. 6246−6253.

[13] Mnih V, Badia AP, Mirza M, Graves A, Lillicrap TP, Harley T, Silver D, Kavukcuoglu K. Asynchronous methods for deep reinforcement learning. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2016. 1928−1937.

[14] Bertsekas DP. Distributed asynchronous computation of fixed points. Mathematical Programming, 1983,27(1):107−120.

[15] Moore AW, Atkeson CG. Prioritized sweeping: Reinforcement learning with less data and less time. Machine Learning, 1993,13: 103−130.

[16] Pan ZY, Zhang ZZ, Chen ZX. Asynchronous value iteration network. In: Proc. of the Int'l Conf. on Neural Information Processing (ICONIP). 2018. 169−180.

[17] Broumi S, Talea M, Bakali A, Smarandache F. Application of Dijkstra algorithm for solving interval valued neutrosophic shortest path problem. In: Proc. of the Symp. Series on Computational Intelligence (SSCI). 2016. 1−6.

[18] Zhang ZZ, Pan ZY, Kochenderfer MJ. Weighted double Q-learning. In: Proc. of the Int'l Joint Conf. on Artificial Intelligence (IJCAI). 2017. 3455−3461.

[19] Sutton RS, Barto AG. Reinforcement Learning: An Introduction. 2nd ed., MIT Press, 2018.

[20] Krose BJA. Learning from delayed rewards. Robotics and Autonomous Systems, 1995,15(4):233−235.

[21] van Hasselt H. Double Q-learning. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). 2016. 2613−2621.

[22] van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proc. of the AAAI Conf. on Artificial Intelligence (AAAI). 2016. 2094−2100.

[23] Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. In: Proc. of the Advances in Neural Information Processing Systems (NIPS). 2016. 3837−3845.

[24] Niepert M, Ahmed M, Kutzkov K. Learning convolutional neural networks for graphs. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2016. 2014−2023.

[25] Franceschi L, Niepert M, Pontil M, He X. Learning discrete structures for graph neural networks. In: Proc. of the Int'l Conf. on Machine Learning (ICML). 2019. 1972−1982.

附中文参考文献:

[1] 孙志军,薛磊,许阳明,王正,深度学习研究综述.计算机应用研究,2012,29(8):2806−2810.

[2] 刘全,翟建伟,章宗长,钟珊,周倩,章鹏,徐进.深度强化学习综述.计算机学报,2018,41(1):1−27.

陈子璇(1996－),女,博士生,CCF 学生会员,主要研究领域为强化学习,智能规划.

潘致远(1993－),男,硕士,主要研究领域为强化学习.

章宗长(1985－),男,博士,副教授,CCF 高级会员,主要研究领域为强化学习,智能规划,多智能体系统.

张琳婧(1995－),女,硕士,主要研究领域为强化学习.