引导式教学场景下深度强化学习的模型研究

汤 胤, 王 雯, 黄书强
(暨南大学管理学院, 广东 广州 510632)

摘要: 针对引导式场景, 结合认知科学上学习区的概念, 构造题库网络图, 进而根据特定学习者的行为来划分割集, 由此建立引导式教学场景下深度强化学习的模型, 在推荐偏差指标的控制下, 做出最适合学习者的内容推荐. 对比实验证明了模型相比控制组能给出合理的"前向推荐", 有效解决学习者作答正确率不稳定的问题. 引导式教学场景下深度强化学习的模型能够拟合经验教师出题决策的思维方式, 在历史作答数据中提取有效隐含信息, 为学习者推荐最佳习题. 模型亦可广泛应用在类似的引导式场景下.

关键词: 推荐算法; 深度强化学习; 学习区; 复杂网络
中图分类号: TP273 **文献标识码**: A **文章编号**: 1000−5781(2020)02−0145−08
doi: 10.13383/j.cnki.jse.2020.02.001

Deep reinforcement learning model in heuristic coaching scenario

Tang Yin, Wang Wen, Huang Shuqiang
(School of Management, Jinan University, Guangzhou 510632, China)

Abstract: By combining the concept of the learning zone in cognitive science, this paper constructs a network of question-base, where the cut set is made based on the behaviors of specific learners, for heuristic scenarios. A deep reinforcement learning model to make the best recommendation for learners is then proposed. The model is trained with the learner's behavior under the control of the recommendation deviation factor, to export the best recommendation of content. A comparative experiment proves the model can effectively solve the unstable problem of the correct rate and outperforms the control group. The model imitates the thinking pattern of an experienced teacher, extracts valid implicit information in the historical answering data, and recommends the best exercises for the learner. The model can also be widely used in similar guidance scenarios.

Key words: recommendation algorithm; deep reinforcement learning; stretch zone; complex networks

1 引 言

传统认知学习理论中非常关注学习者多达 12 个维度的认知风格以及多元智能[1,2], 这对于在线智能体来说意味着庞大的状态空间输入, 也是传统机器学习方法无法提取的隐含的认知特质. 正因为如此, 当前的在线人工智能教育实践中, 习题推荐系统往往存在推荐效率不高, 作答率不稳定, 题目太难或太容易从而让学习者更加容易放弃等各种问题.

近年来兴起的强化学习擅长迭代地适应环境及状态, 典型应用在对抗式场景, 即追求最大化对抗优势从而尽早获得胜利, 例如迅速找到路线或策略、站稳、走出迷宫和击败对手等[3,4]. 对于非对抗式场景, 如自动

收稿日期: 2018−04−10; 修订日期: 2019−01−03.
基金项目: 广东省应用型科技研发专项资金资助项目(2016B010124008); 国家自然科学基金资助项目(71771104); 广州市产学研协同创新重大专项资助项目(201802010034).

驾驶问题[5], 处理的是环境状态的随机性输入, 没有一个难度递进的要求. 这不符合认知科学的范式. 美国心理学家 Tichy 等[6]提出的行为改变理论认为, 最佳的学习方式应当处在一个压力略高于普通水平的学习区, 即舒适区与恐慌区之间. 用于指导学习的人工智能与上述对抗式人工智能应用以及随机性输入的最大不同在于智能体好比一个经验丰富的教师, 必须严格考虑学习者的学习区[7], 不断引导学习者随着学习区内容的掌握而给出略微偏难的问题, 并迭代地向外拓展, 本文称为“ 前向推荐”. 显然, 根据特定学习者行为在庞大题库中不断迭代更新其学习区, 并识别其隐含的认知风格做出最佳的题目推荐, 这是所有面向个性化学习的在线教育都需要面对的问题, 在个性化学习的人工智能探索中往往被忽略. 应当说, 现有的强化学习范式已经较好地解决对抗式场景问题, 但当前研究较少考虑非对抗式场景典型(如引导式教学)的问题, 本文正是在这样的背景下提出的.

引导式场景与个性化推荐也不完全相同. 总的来说目前个性化学习推荐系统主要包括基于学习者及学习资源静态特征和基于学习者动态行为特征两大类. 根据静态特征进行推荐的研究, 如基于内容过滤的推荐系统, 根据学习者需求和学习资源的相似度进行推荐, 导致只会推荐用户已经学习过并有兴趣的资源[8,9], 无法实现“ 前向推荐”. 基于协同过滤进行推荐的研究, 主要依据学习者的评价对相同兴趣点的内容来推荐, 会面临稀疏性和冷启动的问题, 不易发现学习者的新需求[10]. 混合两者的推荐系统也面临推荐精度不高的情况. 也有研究尝试通过社交网络发现共同的兴趣点从而实现推荐, 但这种社会化推荐更容易将陷入认知陷阱, 显然无法满足精准推荐并学习的需求[11]. 同时通过获取静态特征进行推荐的系统往往忽略了学习者行为中蕴含的学习习性等隐含信息. 采用数据挖掘算法能帮助推荐系统获取学习者部分动态行为特征. 如结合使用聚类及关联规则算法得到学习者的学习顺序偏好[12].

因此“前向推荐”这个问题开始受到关注. 在编程学习领域, 序列模式挖掘算法能得到高效学习路径进行推荐学习[13]. 寻找学习者的最佳学习路径还可以用蚁群优化方法, 推送的学习资源有助于提高学习效率与质量[14]. 结合聚类和机器学习技术, 基于学习者特征的相似性度量使用 LSTM 模型来预测学习者的学习路径与绩效[15]. 总的来说, 目前基于深度学习的推荐系统能够使用深度学习模型学习到用户和资源的隐表示, 结合这两种隐表示, 通过内积, Softmax 和相似度计算等方法[16], 做出符合用户兴趣的推荐, 但仍然难以智能地结合用户和资源的隐表示, 也未能拥有教师经验学习的能力, 无法定位学习者的学习区并引导学习者循序渐进挑战难题.

深度强化学习(deep reinforcement learning, DRL)在强化学习基础上引入深度神经网络实现高维函数逼近, 有望解决前文所提到的高维状态空间以及用户行为隐含习性学习的问题, 在类似引导式教学场景的用户交互等领域的表现已经接近或超过了人类水平[17]. 因此本文对于学习区选择策略这一复杂决策问题, 引入深度强化学习模拟有经验教师的决策思维方式, 基于历史作答数据中提取有效学习者认知风格等隐含信息, 就可以面向多元智力和不同认知风格的学生群体引导式地给出最适合学习者的内容, 从而解决上述问题. 研究的难点在于, 智能体对于用户行为建模有较高要求, 同时学习区的具体形式化表达仍然是个难题.

本文对题库搭建网络模型, 从而将寻找学习区的问题转化为定位网络割集的问题. 进而, 利用学习者的作答数据建模, 作为状态输入, 采用深度强化学习算法进行训练, 使得算法能给出准确的割集, 重复实验得到结果进行分析结论.

2 候选题集合获取模型

2.1 题库网络化

为反映题目之间的关联关系, 本文首先对题库搭建网络. 类似人类联想式记忆, 题目网络图的构建分为两步, 首先依据知识点间的相互关系及关系程度构建有向无环加权知识点网络图, 再根据知识点之间是否存在关联及关联程度对同知识点下的题目构建有向无环加权题目网络图.

定义 1 设知识点 a, b 间存在连接, 即 a 为作答 b 前必须掌握的知识点, 则称 a 为前驱知识点, b 为后

继知识点, 记为 $a \to b$.

设知识点网络为 $G(K,E)$, 其中 $K=\{k_i|i=1,2,\ldots,n\}$是网络的知识点节点集合, 节点属性为知识点难度, $E=\{(k_i,k_j)|i,j=1,2,\ldots,n\}$ 是网络中边的集合, 即存在 $k_i \to k_j$ 为先后知识点, 边的关联权重值依据知识点间先后关联程度定义.

类似的, 设题目网络图为 $T=(Q,A)$, 其中 $Q=\{q_i|i=1,2,\ldots,n\}$为题目节点集合, 节点属性由题目在知识点内的难度值确定, $A=\{(q_i,q_j)|i,j=1,2,\ldots,n\}$为边的集合, 边的权重定义为$(q_i,q_j)$ 的关联程度.

进一步地, 对知识点 k_i 对应的题目集合记为 $Q_i=\{q_i|i=1,2,\ldots,n\}\subseteq Q$. 与 k_i,k_j 分别对应的题目是 $p\in Q_i, q\in Q_j$. 若 $k_i\to k_j$, 定义边的集合 $E_{ij}=\{(p,q)|p\in Q_i,q\in Q_j\}$, 则 E_{ij} 边的权重定义为(p,q)的相关性程度; 特别地, 当 $i=j$ 时, p,q 为同一知识点下的题目, p,q 的指向及边的权重依据(p,q)的相关性程度, 由专家指定.

这样, 特定学习者对于习题解答的历史行为, 在图 G,T 中反映为若干成功和失败的节点集合和边, 记为子图 g. 通过 g 可以在图 T 中获得待选题目图 g'(见下节), 基于 g' 可以得到它的一个割集. 由于这个割集的内容处于学习者已掌握内容区域的边缘, 难度水平略高于当前水平, 符合 Tichy 对学习区的定义. 由此, 针对特定学习者的行为寻找学习区问题, 转化为如何在题库网络中寻找特定割集的问题. 形式化表示为

$$f_T(B)\longrightarrow C,$$

其中 T 为题目网络图, C 为学习者割集, B 为学习者历史行为序列. 本文的任务就是在子图 $g\subset T$ 约束下求解 f, 通过学习者历史行为序列 B 得到最佳割集 C.

2.2　在题目网络图中计算候选题集合

记学习者子图 g 中答对的题目为 t_r, 在图 T 中由它指向的节点集合$\{t_b|b=1,2,\ldots,n\}$. 那么已作答题目与它指向的节点连边集合为 $E_{rb}=\{(t_r,t_b)|b=1,2,\ldots,n)\}$. 记子图 g 中答错的题目为 t_w, 在图 T 中找到指向它的节点集合$\{t_f|b=1,2,\ldots,n\}$, 那么已作答题目与它指向的节点间连边集合为 $E_{fw}=\{(t_f,t_w)|f=1,2,\ldots,n)\}$, 令待选题目图 $g'=E_{rb}\cup E_{fw}$. 这个寻找前后向节点的过程见图 1.

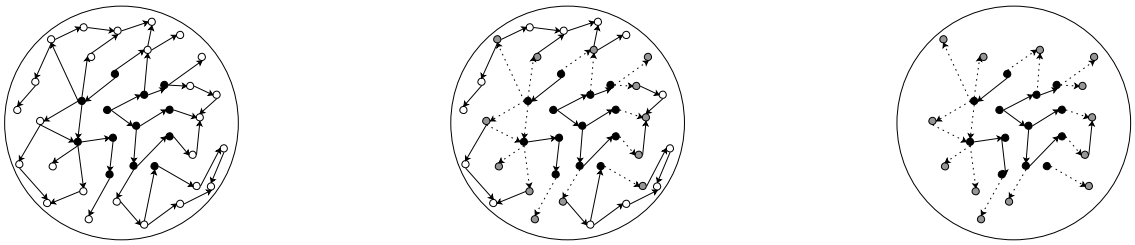

图 1　获取待选题目图的过程

Fig. 1　The process of obtaining a diagram of a topic to be selected

基于待选题目图 g' 中存在多个已作答题目节点指向同一个未作答题目节点的情况, 且题目间的连边权重大小不同, 指的是某一题目在待选题目图 g' 中对于是否下次作答的重要度大小. 如何在未作答题目集合中选择最适合的题目, 显然与连边权重有直接关系. 有向加权网络节点的重要性评估算法将网络中的每条边看作信息流, 结合相应复杂网络的结构特点和“信息量”的定义方法, 求出的节点信息量作为节点的重要性评估指标, 能有效细致地刻画有向加权网络节点之间的差异性[18]. 因而在有向加权的待选题目图 g' 中, 使用该算法得到 g' 中题目节点的重要度排序. 根据 g' 的拓扑结构, 计算 g' 中节点的重要度, 从而对 g' 中节点的重要度属性赋值. 为了不使模型陷入局部最优, 将 t_r 从图 g' 的节点中剔除, 图 g' 剩余节点集合记为候选集合 M.

2.3 推荐偏差度量指标

根据学习区的定义, 理想状态下算法应能每次从学习区给出难度适中的习题, 反映为答对率最终收敛于一个指定的数值. 设学习者每次作答题目数 N, 每次答对个数 n, 构造偏差度量指标为 $d=(n/N-\delta)^2$, 其中 $\delta\in[0,1]$, $n\leqslant N$. d 表示学习者单次得分与我们期望的偏差, 显然学习者作答的题目越是处于学习者的学习区, d 越接近 0. 考虑到算法运行次数及稳定性, 实际实验中考察的是每 100 回合(epoch)或指定区段中推荐偏差的均值, 即 $e=\dfrac{1}{k}\sum\limits_{i=1}^{k}d_i$, 其中 k 为该 100 回合的作答次数.

3 引导式教学场景下深度强化学习的模型

在联想式题库基础之上, 使用 DQN[4]算法, 结合经验池回放技术, 构建算法如下:

步骤 1 初始化经验池及神经网络. 经验池 D(R)用于存储训练样本, R 为容量大小; action-value 函数 Q, 权重参数 $\theta\leftarrow$ random( ); 目标 action-value 函数 $\hat{Q}$, 权重参数 $\theta'\leftarrow\theta$; 总回合次数 E;

步骤 2 获取学习者状态 s, 根据 s 在题目网络图 T 中计算待选题目图 $g'=E_{rb}\cup E_{fw}\subset T$;

步骤 3 剔除图 g' 中状态 s 作答正确的节点, 计算候选集合 $M=t_{g'}-(t_{g'}\cap t_r)$;

步骤 4 对重要度属性排序;

步骤 5 在候选集合 M 中选择 N 题, 以概率 $1-\varepsilon$ 选择 $a_t=\arg\max\limits_{a}Q(\phi(s_1),a;\theta)$, 其余随机选择;

步骤 6 获取用户作答数据, 计算奖励 $r_t=\begin{cases}p, & n/N=\delta\\ -|n-\delta\times N|, & n/N\neq\delta\\ d, & \text{答题回合结束;}\end{cases}$

步骤 7 读取下一训练样本 x_{t+1}, 更新状态 $s_{t+1}\leftarrow s_t,a_t,x_{t+1}$;

步骤 8 将本次运行经验(s_t,a_t,r_t,s_{t+1})存入经验池 D;

步骤 9 随机选取$(s_j,a_j,r_j,s_{j+1})\subset D$, 共 minibatch 条;

步骤 10 若回合结束, $y_i\leftarrow r_j$, 否则 $y_i\leftarrow r_j+\gamma\max\limits_{a'}\hat{Q}(s_{j+1},a';\theta')$;

步骤 11 每隔 C 次迭代计算 $\theta\leftarrow\theta+\Delta(y_i-Q(s_j,a_j;\theta))^2$, $\theta'\leftarrow\theta$;

步骤 12 若未达到总回合次数 E, 重复步骤 2～ 步骤 11; 否则退出循环并输出.

3.1 动作集合

从候选集合 M 中选题的策略包括根据题目节点的综合难度值和根据重要度值两种, 共 8 个动作(表 1), 其中 $j+h+l=N$, N 为每次给学习者作答选择的题目数量.

表 1 选题动作

Table 1 Actions of question selection

编号(No.)	动作(a)
1	综合难度最高 N 题
2	综合难度适中 N 题
3	综合难度最低 N 题
4	综合难度最高 j 题, 适中 h 题, 最低 l 题
5	重要度最高 N 题
6	重要度适中 N 题
7	重要度最低 N 题
8	重要度最高 j 题, 适中 h 题, 最低 l 题

3.2 环境状态表示

环境状态包含学习者的状态和作答对错的反馈, 考虑到模型的效率, 以学习者最近 i 题的作答情况作为

初始状态, 进而根据新的作答情况更新状态, 将学习者在此前i题的作答情况用一个一维数组 s_t 表达, 即

$$s_t = \{x_1, x_2, \ldots, x_i, 1, 0, \ldots, 1\},$$

其中 x_i 表示学习者作答的题目信息, 0 和 1 为作答情况, 0 表示答错, 1 表示答对.

3.3　即时策略奖励

模型的目标是给出对于特定学习者来说难度适中的习题, 学习者对选出 N 题的作答正确数为 n, 则根据 n 给出反馈奖励 r, 如表 2 所示. 答对率满足度量指标 δ 值的回报值为 p, d 为答题回合结束的回报值.

表 2　出题策略的奖励规则

Table 2　Reward policy

Event	Reward(r)
$n/N = \delta$	p
$n/N \neq \delta$	$-\lvert n - \delta N\rvert$
答题回合结束	d

4　实验设计及结果分析

4.1　实验数据

实验题目数据来源为游戏化学习软件“等于”的后台数据库, 参考小学数学课程老师的意见, 对小学数学四年级上学期的 1 479 道题目构建有向加权无环题目网络图, 形成 275 476 条连边. 鉴于本文算法所使用的题目信息、用户行为信息的特殊性、推荐粒度以及应用领域的不同, 无法与经典个性化学习推荐系统在推荐效率及算法层面进行直接比较. 因此参考文献[14]的做法, 直接从学习者学习效率角度, 设置实验组和控制组进行对比. 共 160 名小学四年级学生参与实验, 按照性别、年龄、数学成绩进行随机分为两组, 确保两组同质性. 控制组在完成一轮习题后, 采用随机方式选题, 与无经验的教师类似. 相比控制组, 实验组则使用本文算法选题. 共获取实验组 80 个用户的 11 832 条行为数据, 控制组 80 个用户 11 832 条行为数据, 每个用户数据重复训练 20 000 回合. 实验的主要目的是观察推荐偏离指标 e 是否能够显著低于控制组.

4.2　实验参数及流程设计

模型参数设置及训练超参设置分别如表 3 和表 4 所示, 设定当选择的 N 题全为该学期最后一个知识点时回合结束.

表 3　模型参数设置

Table 3　Parameters for Model

参数	大小	描述
i	100	取当前状态前 i 题作答情况
x_i	综合难度值	前 i 道题目的综合难度值
N	10	每次选题的数量
j	3	排序最高的题目数量
h	4	排序居中的题目数量
l	3	排序最低的题目数量
δ	0.8	最优推荐指标
p	3	$d = 0$ 的回报值
d	0	回合结束回报值

表 4　训练超参设置

Table 4　Hyperparameters for Training

参数	大小	描述
E	20 000	每个用户实验回合数
minibatch	32	每次从经验池中回放抽取的数量
R	500	经验池容量
折扣因子 γ	0.9	Q 值更新时的折扣因子
ε-greedy	0.9	ε-贪心探索的 ε 大小
学习率	0.01	参数更新的学习率
C	200	目标网络的参数更新间隔
激活函数	ReLU	网络中的激活函数
隐藏层	2	网络中隐藏层的数目
输入层神经元数	20	输入层的神经元数量
隐藏层神经元数	20	每层隐藏层的神经元数量
输出层神经元数	8	输出层的神经元数量

4.3　结果分析

为了观察每次模型选题策略的决策情况, 考察推荐偏差指标 e 的变化. 图 2 和图 3 分别为用户每 100 回

合推荐偏差的均值和方差. 控制组(control)维持在一个比较高平均值和方差, 符合人们常规对无经验教学的印象, 低于人类表现. 而实验组(experimental)在训练初期均值及方差都较大, 随着训练推进, 均值及方差都降低并趋于稳定, 意味着模型在训练中能有效学习出题策略, 并且从特定用户的学习区中出题的准确性随着训练次数的增加不断地提高且趋于稳定, 也超出了人类表现. 图中 baseline 指的是人类表现, 由有经验教师重复多次考察得到.

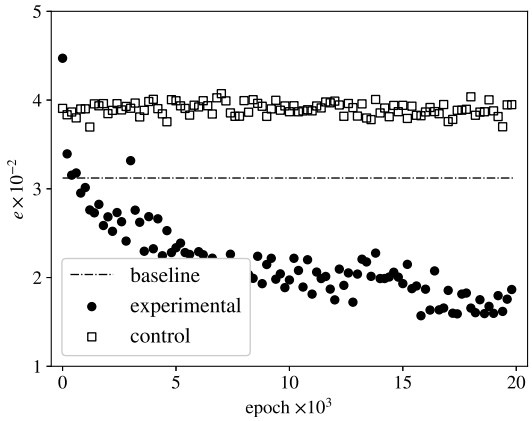

图 2 推荐偏差的均值

Fig. 2 Mean value of recommendation deviation

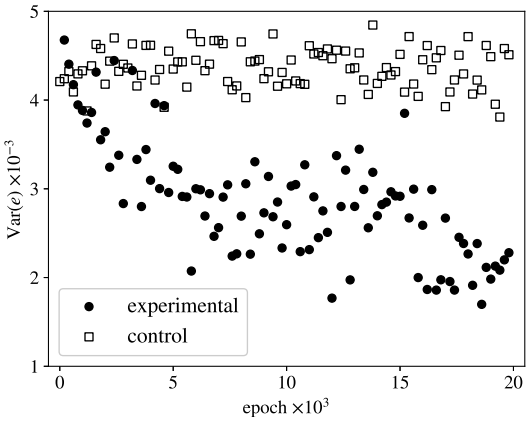

图 3 推荐偏差的方差

Fig. 3 Variance of recommendation deviation

为了观察算法的稳定性, 实验组获取学习者每 100 回合中平均作答的得分频率分布 f(图 4), 即每 100 回合中学习者作答正确数的分布情况. 随着回合数的增加, 从 100∼200, 201∼300, 301∼400 回合的得分频率分布曲线可以明显看出学习者的作答正确数 n 有逐渐向 $\delta N = 8$ 聚集的趋势. 在训练的最后 100 回合, 学习者每次作答正确数为 δN 的次数也显著增加, 并且作答正确数主要分布在 $\delta N = 8$ 的周围. 作为对照, 控制组(图 5)在同样的得分频率曲线中, 观察不到 n 收敛于 $\delta N = 8$ 的趋势.

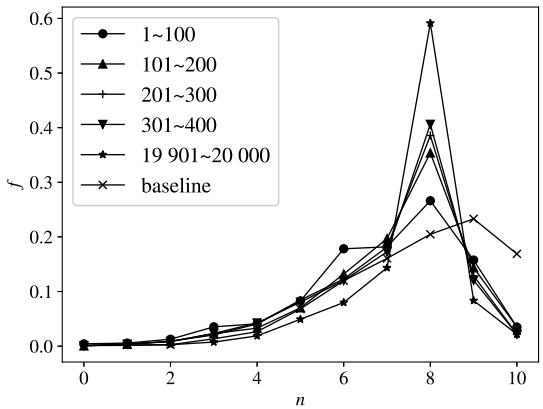

图 4 实验组每100回合中作答正确数分布

Fig. 4 Correctness distribution in every 100 epochs (Experimental group)

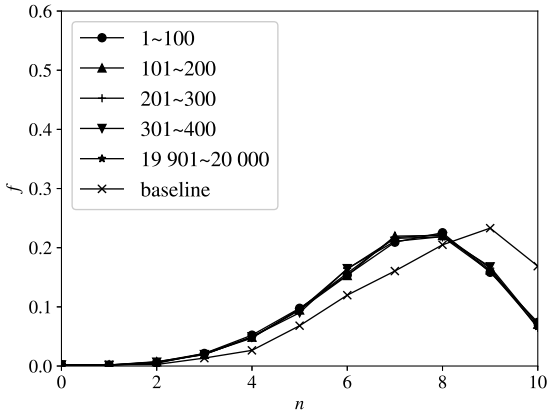

图 5 控制组每100回合中作答正确数分布

Fig. 5 Correctness distribution in every 100 epochs (Control group)

从另外一个角度观察深度强化学习的学习过程, 考察动作的价值函数即 q 值在训练过程中是否稳定收敛. 图中定义 $q = \max\limits_a Q(s,a)$, 代表模型对某状态下使用最佳动作能够获得的期望折扣回报值. 模型对 ε-greedy 的值设定为 0.9, 即在每次选动作时会有 0.1 的概率随机选择动作. 为了能准确评估 q 值的变化,

筛选出每回合初始状态时都采取最佳动作时的 q 值共 18 047 个. 图 6 表明, q 值在前期的训练中有所震荡, 说明模型在探索策略过程中, 训练的中后期逐渐趋于稳定, 说明模型得到了训练并取得了较好的学习效果.

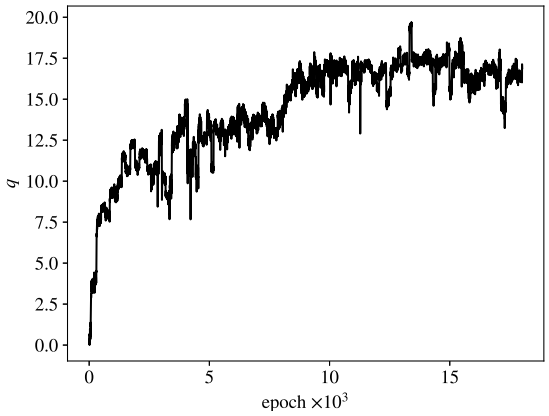

图 6. 每回合初始状态的价值函数

Fig. 6　The value function of the initial state of each epoch

5　结束语

本文解决了传统基于数据挖掘的习题推荐模型难以处理的“前向推荐”和“隐含认知风格获取”问题, 一方面通过在深度神经网络的隐含特征提取能力, 对不同学习者的风格和多元智能特性均可较好适应并做出不同的推荐, 另一方面, 由于考虑了知识点之间及题目之间的相关性这个重要特征, 作答者的答对率保持稳定, 意味着其智能体给出的推荐收敛在了作答者的学习区. 本文的贡献在于, 考虑学习者学习能力水平的复杂性和随着学习过程导致的易变性, 针对引导式场景, 在学习者行为数据基础上提出了一种基于引导式教学场景下深度强化学习模型的算法: 首先基于学习者的作答历史数据信息找到学习者在题目网络图中的待选题目图, 然后结合深度强化学习的经验池回放技术及在线学习方法对网络模型进行训练. 实验结果显示算法的学习效率明显超越控制组, 学习者的作答得分值及推荐质量趋于稳定, 证明了模型的有效性.

本文的局限性在于, 由于小学生用户配合的难度以及数学学科的特点, 决定了数据选择的质量, 在其他学科有可能需要做一定调整. 另外一方面由于考虑了题目之间的相关关系, 搭建的题目网络一方面为算法提供了智能选择的依据, 同时也限制了其在其他领域的应用. 今后的工作可以在多方面展开, 如实验方面扩充到其他学科和其他年级, 进一步验证算法的适用性. 另外本文目前基于经典的 DQN 算法, 事实上在研究过程中更高效的强化学习算法典型如 A3C, DDPG, Ape-X, Rainbow 层出不穷, 今后也考虑在保证学习效果的基础上更新算法.

参考文献:

[1] Coffield F J, Moseley D V, Hall E, et al. Should We Be Using Learning Styles: What Research Has to Say to Practice. London: Learning and Skills Research Centre, 2004: 84.

[2] Gardner H E. Multiple intelligences: New horizons in theory and practice. Circulation, 2006, 96(10): 3647–3654.

[3] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with Deep Reinforcement Learning. Lake Tahoe: NIPS Deep Learning Workshop, 2013: 231.

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

[5] Sallab A E, Abdou M, Perot E, et al. Deep reinforcement learning framework for autonomous driving. Electronic Imaging, 2017, 2017(19): 70–76.

[6] Tichy N M, Sherman S. Control Your Destiny or Someone Else Will. New York: HarperBus, 2005.

[7] Ericsson K A, Charness N, Feltovich P J, et al. The Cambridge handbook of expertise and expert performance. Journal of Workplace Learning, 2006, 20(7): 560–560.

[8] Wang Z, Liu Y, Yang J, et al. A personalization-oriented academic literature recommendation method. Data Science Journal, 2015, 14: 17

[9] Salehi M, Kamalabadi I N, Ghoushchi M B G. Personalized recommendation of learning material using sequential pattern mining and attribute based collaborative filtering. Education and Information Technologies, 2014, 19(4): 713–735.

[10] Niu J, Wang L, Liu X, et al. FUIR: Fusing user and item information to deal with data sparsity by using side information in recommendation systems. Journal of Network and Computer Applications, 2016, 70: 41–50.

[11] Seo Y D, Kim Y G, Lee E, et al. Personalized recommender system based on friendship strength in social network services. Expert Systems with Applications, 2017, 69: 135–148.

[12] Aher S B, Lobo L M R J. Combination of machine learning algorithms for recommendation of courses in e-learning system based on historical data. Knowledge-Based Systems, 2013, 51(1): 1–14.

[13] Nakamura S, Nozaki K, Nakayama H, et al. Sequential pattern mining system for analysis of programming learning history // IEEE International Conference on Data Science and Data Intensive Systems. 2015: 69–74.

[14] Eugenijus K. Recommending suitable learning paths according to learners' preferences: experimental research results. Computers in Human Behavior, 2015(51): 945–951.

[15] Zhou Y, Huang C, Hu Q, et al. Personalized learning full-path recommendation model based on LSTM neural networks. Information Sciences, 2018, 444: 135–152.

[16] 黄立威, 江碧涛, 吕守业, 等. 基于深度学习的推荐系统研究综述. 计算机学报, 2018, 41(7): 1–29.
Huang L W, Jiang B T, Lü S Y, et al. Survey on deep learning based recommender systems. Chinese Journal of Computers. 2018, 41(7): 1–29

[17] Dulac-Arnold G, Evans R, Van Hasselt H, et al. Deep reinforcement learning in large discrete action spaces. https://arxiv.org/abs/1512.07679, Apr. 2016.

[18] 王 班, 马润年, 王 刚, 等. 改进的加权网络节点重要性评估的互信息方法. 计算机应用, 2015, 35(7): 1820–1823.
Wang B, Ma R N, Wang G, et al. Improved evaluation method for node importance based on mutual information in weighted networks. Journal of Computer Applications, 2015, 35(7): 1820–1823.

作者简介:

汤 胤(1975—), 男, 福建周宁人, 博士, 副教授, 研究方向: 大数据与商务智能, Email: ytang@jnu.edu.cn;

王 雯(1993—), 女, 安徽亳州人, 硕士, 研究方向: 大数据与商务智能, Email: 526471887@qq.com;

黄书强(1975—), 男, 湖北麻城人, 博士, 教授, 研究方向: 认知无线网络, 物联网, 计算智能, Email: hsq@jnu.edu.cn.