基于强化学习的作战辅助决策建模研究*

朱宁龙[1]　佟骁冶[2]
（1. 91404部队92分队　秦皇岛　066000）（2. 中国船舶集团有限公司第七〇九研究所　武汉　430205）

摘　要　传统的辅助决策方法中，专家系统通常由固定知识表达，完全依赖领域专家知识，不具备学习能力。为了弥补专家系统的局限性，利用对智能体进行训练的方法自动生成决策方案。分析了自动机和行为树等过程性建模方法的优缺点，提出了基于Petri网和Q学习的CGF行为与决策建模方法。介绍了模型的基本结构和强化学习机制。在虚拟场景中，设计了决策模型实现方法和学习规则，通过对比实验，证明了此方法的可行性。

关键词　CGF；Petri网；强化学习；行为；决策
中图分类号　TP273　　**DOI**：10. 3969/j. issn. 1672-9730. 2021. 07. 009

Research on Modeling of Operational Assistant Decision Based on Reinforcement Learn

ZHU Ninglong[1]　TONG Xiaoye[2]
（1. Unit 92，No. 91404 Troops of PLA，Qinhuangdao　066000）
（2. 709th Research Institute，China State Shipbuilding Co.，Ltd.，Wuhan　430205）

Abstract　In the traditional assistant decision-making methods，the expert system is usually expressed by fixed knowledge，which completely depends on the domain expert knowledge and has no self-learning ability. In order to make up for the limitation of expert system，the method of training agent is used to generate decision scheme automatically. After analyzing the advantages and disadvantages of process modeling methods such as automata and behavior tree，a CGF behavior and decision modeling method based on Petri net and Q-learning is proposed. The basic structure of the model and reinforcement learning mechanism are introduced. In the virtual scene，the implementation method and learning rules of decision model are designed. The feasibility and good performance of this method are proved by comparative experiments.

Key Words　CGF，Petri net，reinforcement learning，behavior，policy
Class Number　TP273

1　引言

目前，指控系统中的辅助决策知识主要通过人工制定，这种方法低效而又充满主观性。在大数据和智能化背景下，可以利用虚拟环境下的作战数据辅助生成决策，利用CGF（Computer Generated Force，计算机生成兵力）帮助指挥员进行决策训练与分析。传统CGF的行为决策按照预设规则制定，基于“IF…THEN”指令形式的知识，进行产生式推理，智能化程度不够高，而对于军事领域来说，一般的机器学习方法又缺乏可靠性与可解释性。同时具备严谨的知识结构和高效的学习能力，是实现辅助决策智能化生成的关键。

考虑到军事领域的特殊性，在学习方法的构建中，既需要遵循专家领域知识，体现规则的约束力，又要通过自学习，自动寻找最优解。目前有很多研究将状态机等知识表示工具与神经网络等机器学习工具相结合，训练出理想的行为决策[1]。而对于辅助决策场景来说，需要选择具有严谨的知识表示形式的模型构建方法，而不是通过纯粹的概率计算或算法求解实现推理的方法。并且，学习的过程必须有可解释性，才能有可信度，也方便指挥员等对模型的训练结果进行信息提取，分析，得出想要的结论。

* **收稿日期**：2021年1月20日，**修回日期**：2021年2月5日
作者简介：朱宁龙，男，硕士，研究方向：指挥控制系统。佟骁冶，男，硕士，工程师，研究方向：指挥控制系统。

强化学习方法是目前规则学习的重要手段，通过不断交互的方式获取学习数据，从而得出策略。同时，强化学习属于人工智能中的行为主义流派，其学习过程符合人类行为决策的生成过程，从基本原理上具备一定的可解释性。强化学习的学习模式，即“状态空间—行为空间—回报值”，也非常符合人类行为与认知的规律。Rahul Dey 和 Chris Child 的研究[2]使得智能体同时具备固定知识与学习能力，他们将行为树与 Q 学习算法相结合，设计了可以自学习的行为树结构，取得了不错的效果，体现了过程建模和强化学习相结合的思想。另外也有研究将此方法应用于 CGF 行为优化[3]。在此基础上，本文提出了将 CGF 的行为决策知识以 Petri 网的方式存储，并以 Q 学习更新 Petri 网结构的方法，使知识模型具备自学习能力。

2　模型结构与学习方法

决策行为过程具有马尔科夫特性，即当前状态向下一个状态转移的概率和后果只取决于当前状态和当前发生的行为，与之前状态无关。决策行为的自学习可以依据马尔科夫过程分步实现，对当前状态做出反应，然后获取系统的反馈，再得出当前步骤的学习结论。根据这一特性，链状或者网状模型最适合对此类过程进行建模，并通过强化学习的方式进行模型更新[4-5]。

为了解决行为树的本身存在结构不灵活，要素过多需要循环检测机制驱动，造成重复遍历的缺点，本文采用 Petri 网表达模型结构，既解决了自动机面对复杂环境时的状态空间爆炸问题，并且具备了并行过程表达能力。Petri 网对状态机最重要的继承是“状态”这一概念，不需要循环检测机制就可以触发动作，减少了重复遍历。另外，Petri 网中元素种类比较少，只有“库所”、“变迁”两种类型的节点，“增删改”相对行为树更加灵活。比行为树更具有优势的地方在于：Petri 网是一种系统的数学和图形分析工具，可以清楚地描述模型中各个元素的相互关系和动态行为，既具有工作流[6]的特征，又有成熟的数学分析方法。

对于强化学习来说，状态既可以来自于系统变量，也可以来自于模型本身。Petri 网可以读取状态，也可以改变状态。环境状态体现状态空间中的元素，逻辑状态体现决策进行的阶段，分别用两种库所表示，将动作用变迁表示，当前置库所均满足条件，即可触发动作[7-8]。形式如图 1 所示。

在传统 Petri 网中，变迁的触发取决于前置库所中的托肯值和输入弧的权重。为了训练网模型，对传统网模型进行结构改造，赋予变迁额外的参数 Q 值，用来表示当前动作的优先程度。根据 Q 值将平行变迁（拥有相同前置库所的变迁）进行排序，把每轮训练结果中 Q 值最大的变迁赋予最高优先级，如图 2 所示。

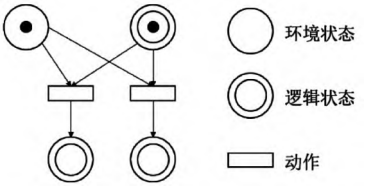

图 1　决策行为表示

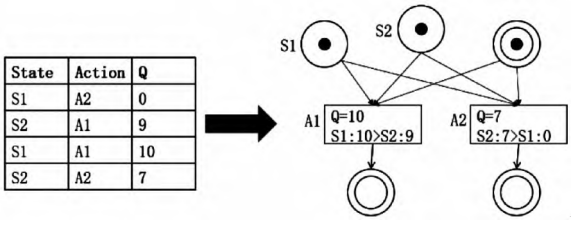

图 2　Petri 网中引入 Q 值

同时，将此最大 Q 值赋予上游变迁，即当前输入库所的输入变迁，使得上级平行变迁获得排序，作为上级决策的执行依据，以此类推，层层传递。Q 值传递方式如图 3 所示。

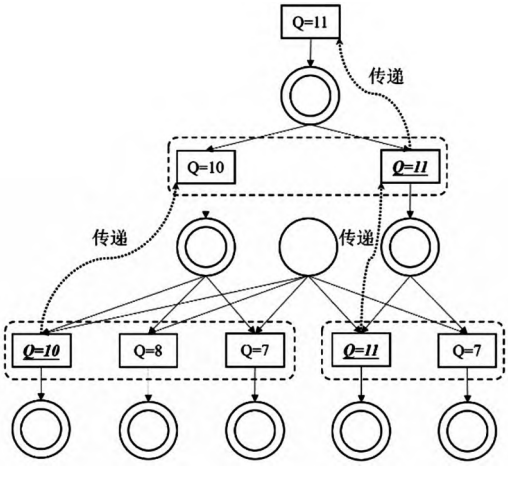

图 3　Q 值的传递

由此方式训练的结果是从初始状态节点开始，根据当前最大 Q 值选择触发候选动作，并以同样的方式选择后续动作节点。训练结束后，剔除低顺位变迁以及相关库所，最终决策方案得以确定。

3　实验设计

在某型指控系统平台中，建立虚拟环境，设置两组智能体（A 组、B 组），对其行为分别以状态机（A 组）和 Petri 网（B 组）的形式进行建模，并在其中一组实验中，在对战过程中对 B 组智能体进行强化

学习训练。

3.1 网模型的表示方法

采用PNML标准制定Petri网模型标记语言,pnml标签表示PNML文件,net标签表示网络,库所、变迁和弧分别用place标签、transition标签和arc标签表示[9~10],其文件基本结构如下所示。

```
<pnml>
<net id="net1">
<page id="page1">
<place id="place1">…</place>
<transition id="transition1">…</transition>
<arc id="arc1" source="place1" target="transiton1">…</arc>
……
</page>
</net>
</pnml>
```

其中,库所的初始托肯值用“initialMarking”属性表示,库所名称用“Pname”属性表示,其结构如下。

```
<place id="place1" initialMarking="1" Pname="状态1">
</place>
```

变迁的名称用“Tname”属性表示,其相应的Q值用“Qvalue”属性表示,其结构如下。

```
<transition id="t1" Tname="动作1" Qvalue="10">
</transition>
```

Petri网中的弧分为两种,即“库所-变迁”和“变迁-库所”,分别以两种方式表示,结构如下。

```
<arc id="a1" source="place1" target="transition1" weight="1">
</arc>
<arc id="a2" source="transition1" target="place2" weight="1">
</arc>
```

PNML表达了网络中各个元素之间的关系,系统运行时,只需通过对PNML文件的读取和操作,生成关联矩阵临时数据和元素实体,并在模型训练过程中修改PNML文件,表达为新的决策模型[11]。

3.2 状态设置

确立了五个状态元素,为了减小状态空间,提升学习速度。对环境中的状态数据进行模糊化和离散化的预处理,选取相关的状态和动作。

1)健康状态H(None,Low,Medium,High);

2)弹药储备A(None,Low,Medium,High);

3)敌人距离D(Near,Medium,Far);

4)受到攻击U(Yes,No)。

3.3 行为设置

Agent的行为并不是单一动作,而是一系列动作组成的高层次行为组合。包括以下四种。

1)巡逻patrol;

2)防御defend;

3)攻击attack;

4)逃避dodge。

由于仿真环境中,行为是持续性的,存在一个运行周期,Petri网的并行特性允许各个行为同步发生,完成实时仿真和实时学习。

3.4 奖励函数设置

首先设置基本奖励函数机制,将战损、被消灭的战果给与惩罚,将重创、击毁的战果给与奖励。为了加快算法的收敛速度,防止从“零”开始学习引起的低效。将专家经验作为先验知识,与基本奖励机制加权后相结合,组成综合奖励函数,综合奖励函数R表达式为

$$R=\omega_1 R_1+\omega_2 R_2$$

其中,ω_1、ω_2分别为先验知识权重和基本奖励函数权重,R_1、R_2分别为先验知识奖励函数和基本奖励函数。

本实验的先验知识奖励函数部分展示在表1中。

表1 先验知识奖励函数

	H	A	D	U	Action	Reward
S1	None	-	-	-	-	-100
S2	Low	Low	Medium	Yes	Dodge	10
S3	High	High	Far	No	Patrol	0
S4	High	High	Near	Y/N	Attack	30
……	……	……				……

3.5 模型训练

Q-Learning算法,是一种异策略控制(Off-policy)的采用时序差分法的强化学习方法,使用两个控制策略,一个用于选择新的动作,另一个用于更新价值函数。

使用bellman方程对马尔科夫过程求最优策略,其算法流程如下:1)初始化$Q(s,a)$;2)根据当前Q值,选择当前状态s下的一个动作a(可使用ε-greedy搜索策略),输出动作a后,观察输出的状态s'和奖励r,依据公式更新Q值:$Q(S,A)=Q(S,A)+\alpha(R+\gamma\max_a Q(S',a)-Q(S,A))$,更新策略:$\pi(a|s)=\arg\max Q_a(s,a)$;3)重复“步骤2)”,直到$Q$值收敛。其中,$\pi$是当前策略,$\gamma$是衰减因子。

为了避免陷入局部最优,兼顾当前最优解之外

的可能性，也为了解决冷启动问题，采用 ε -greedy 算法为动作选取策略。设置一个概率值 ε 、随机数 *rand* 和参数 t ，算法每步都有一定的概率 ε 在可选动作集 A 中选择探索，也有一定的概率 $1-\varepsilon$ 进行采样。搜索策略表示为[12~13]

$$\pi(a|s)=\begin{cases} random(A) & \text{if } rand<t \\ \arg\max_a Q(s,a) & \text{else} \end{cases}$$

学习后期，agent 对 Q 值的学习方向越来越明朗，可以适当减小 ε 值，直至取消搜索。

截取某一逻辑节点的情况下，对于特定状态 S2（健康状态：Low、弹药储备：Low、敌人距离：Medium、受到攻击：Yes），其初始模型如图 4 所示，此情况下智能体随机采取四种策略。

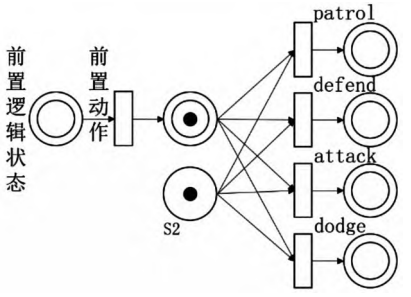

图 4　初始模型

设置学习效率 $\alpha=0.8$ ，折扣因子 $\gamma=0.6$ ，初始概率 $\varepsilon=0.25$。经学习收敛后 4 个动作的 Q 值分别为 Q（patrol）=1、Q（defend）=5、Q（attack）=9、Q（dodge）=15，经过低顺位动作剔除，最终形成的决策模型如图 5 所示。

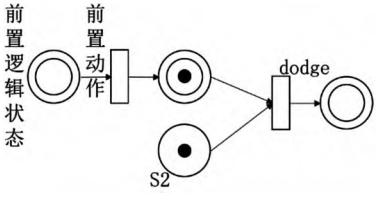

图 5　训练后模型

对所有的逻辑节点下的状态采取同样方式训练，以图 3 所示的方法进行 Q 值传递，最终产生完整的决策网络。

4　实验结果

将 A 组和 B 组设为完全相同的简单的初始决策。令两组智能体进行“5V5 对抗”，记录每一局的存活数。每次实验总局数为 50，每局时间限制在 1min。

实验分为两部分，首先：使未经学习的 B 组与 A 组进行对抗，此时 A、B 两组具有相同的决策，A、B 两组的对抗实验结果如图 6 所示。

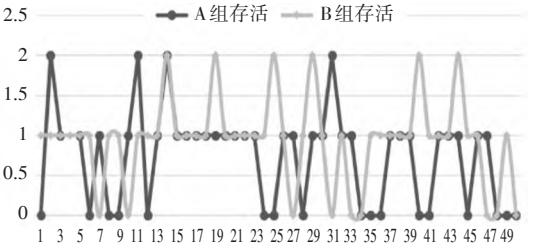

图 6　对战存活单位数量对比图（实验一）

其中，横轴表示对抗次数，竖轴表示存活数量。对抗实验结果表明同样的行为决策下，拥有相同决策的两组智能体的成绩相似。

在第二次实验中，B 组的 Petri 网模型加入 Q 学习更新机制，自第二轮对抗开始，B 组每轮对抗中使用的策略都是上一轮学习后的结果。训练分为 50 个周期，截取每个周期的最后一局结果，实验结果如图 7 所示。

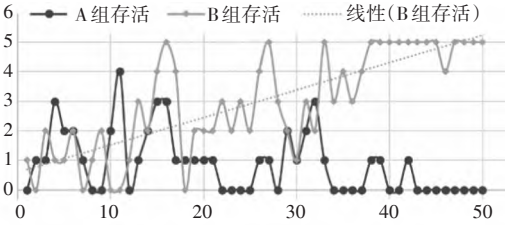

图 7　对战存活单位数量对比图（实验二）

从结果来看，两次实验有明显差异。在第二次实验中，在前期对战，由于 A 组和 B 组都按照相同或相似的策略行动，结果互有胜负，胜负主要取决于随机因素，双方战力比较均衡。前中期，因为随机搜索机制的影响，B 组可能会选择错误的决策行为，战绩不太稳定。但是随着强化学习的进程，搜索范围逐步收窄，B 组战绩开始提升，并且越来越稳定。

实验结果表明，用 Petri 网建模方法的智能体具有与状态机相似的决策能力，经过强化学习的 Petri 网决策模型有一定的成长性，可以一定程度上规避局部最优的困境，并且学习效率比较可观。

5　结语

本文提出的基于 Petri 网建模和强化学习的指控系统辅助决策生成方法，可以由计算机生成兵力训练决策方案，将最优解提供给指挥员。在建模的便捷性、可解释性和成长性方面都取得了不错的效果。使用此方法一定程度上弥补了人为设置决策模型的缺陷，可辅助决策行为知识建模。目前此方法还没有将更复杂的状态空间以及行为空间引入，如何更好地处理模糊类型的变量和连续变量，以及

如何更加合理地动态调节搜索力度,平衡算法运行的速度和容错率,是下一步要研究的工作。

参考文献

[1] 王钦钊,张心路,郭傲兵. Game AI在计算机生成兵力中的应用研究[J].计算机应用研究,2020,37(S1):17-18,5.

[2] Dey R, Child C. QL-BT: Enhancing behaviour tree design and implementation with Q-learning[C]// Computational Intelligence in Games (CIG), 2013:275-282

[3] 付延昌. 基于行为树的CGF行为建模研究[D]. 长沙:国防科学技术大学,2016.

[4] 石轲. 基于马尔可夫决策过程理论的Agent决策问题研究[D]. 合肥:中国科学技术大学, 2010.

[5] 郭靖. 基于马氏决策理论的智能体决策问题研究[D]. 广州:广东工业大学,2012.

[6] 陈慧灵,王宪增,邹宽城. 基于Petri网的工作流过程建模[J].计算机工程与科学,2008(05):92-94,105.

[7] 姜浩,董逸生,罗军舟.基于Petri网的工作流过程建模与分析方法的研究(英文)[J]. Journal of Southeast University(English Edition),2000(02):66-73.

[8] 袁杰,李伟. 运输机器人行为建模的Petri网方法[J].计算机应用,2014,34(05):1360-1363,1368.

[9] 胡晓静,胡敏,刘士喜. Petri网标记语言[J].计算机技术与发展,2011,21(12):66-69.

[10] 吴振寰,王鹏伟. Petri网关联矩阵与PNML描述之间的转换[J].计算机工程与应用,2006(21):32-34.

[11] 宋瑜辉. 基于UML和Petri网建模的研究与应用[D].西安:西安建筑科技大学,2009.

[12] 刘晓伟,高春鸣. 结合行为树与Q-learning优化UT2004中agent行为决策[J].计算机工程与应用,2016,52(03):113-118.

[13] 张汝波,顾国昌,刘照德,等. 强化学习理论、算法及应用[J].控制理论与应用,2000(05):637-642.

(上接第27页)

化的复杂条件下航行时,航海人员对航行环境和安全态势的感知仍然依赖于导航信息在ECDIS上的叠加显示、查阅航海图书资料和预报图表以及传统的航迹绘算和海图作业,然后以人工决策的方式对航行环境变化进行分析。人工查阅或计算多种环境要素资料、数据的过程繁琐,甚至可能因为航海人员缺乏经验或疲劳、疏忽造成错误,使得航行环境威胁态势信息的实时性、精准性和可视性不高。如果将舰船航行环境威胁态势及其变化作为一种地理对象进行信息建模和显式表达,有望突破传统ECDIS隐式表达航行环境威胁态势的局限,实现向直接表达动态航行环境为主并显式描述威胁态势变化机制的转变。

参考文献

[1] 张安民. e-航海中的动态信息服务若干关键技术研究[D]. 武汉:武汉大学博士学位论文,2013:44-65.

[2] 张立华. 基于电子海图的航线自动生成理论与方法[M]. 北京:科学出版社,2011:55-60.

[3] 王代锋, 洪华生. 海基于潮汐表数据同化的天文潮数值预报模型及其模拟预报效果[J]. 台湾海峡,2010,29(2):154-158.

[4] 张安民,杜佳芸,王蕊,等. e-航海架构的高精度动态水深服务实现[J]. 测绘科学,2018,43(7):149-155.

[5] 江应境,高山红. 一种动态权重的台风集成预报方法[J]. 海岸工程,2018,37(3):1-13.

[6] 高珊,朱翊,张福浩. 基于GIS的台风案例推理模型[J]. 测绘科学,2013,38(6):46-48.

[7] 马娟娟,孙海燕. 基于GIS的台风预警系统设计与实现[J]. 地理空间信息,2014,12(1):134-136.

[8] 王中山,袁金锦. 面向灾害风险评估的台风数值模拟及可视化[J]. 测绘通报,2015(4):108-110.

[9] 张进峰,王晓鸥,刘永森. 基于动态风浪环境的我国近海船舶避台航线优化[J]. 中国航海,2016,39(2):45-49.

[10] 朱庆,谭笑,谢林甫,等. 机场环境威胁态势信息在语义空间的统一建模及其导航应用[J]. 武汉大学学报(信息科学版),2015,40(3):341-346.

[11] 冯文娟,杜云艳,苏奋振. 台风时空过程的网络动态分析技术与示例[J]. 地球信息科学,2007,9(5):57-63.

[12] 季民. 海洋渔业GIS时空数据组织与分析[D]. 青岛:山东科技大学博士论文,2004:45-51.

[13]刘文亮,苏奋振,杜云艳. 海洋标量场时空过程远程动态可视化服务研究[J]. 地球信息科学学报,2009,11(4):513-519.

[14] 薛存金,董庆. 海洋时空过程数据模型及其原型系统构建研究[J]. 海洋通报,2012,31(6):667-674.

[15] 龚健雅,李小龙,吴华意. 实时GIS时空数据模型[J]. 测绘学报,2014,43(3):226-232.