* 1

基于多 Agent 强化学习的战时备件供应保障动态协调机制

刘喜春，　王　超，　王文广，　王维平
(国防科技大学　五院系统工程系，湖南　长沙　410073)

摘　要：有效的备件供应保障是保证航空装备处于良好状态的重要因素。战时备件供应保障的不确定性特点突出。为了应对这些不确定，精确保障要求下战时备件供应保障更加强调备件资源在系统中的动态协调。鉴于战时备件供应保障系统与多 Agent 系统的相似性，采用基于 Agent 的建模仿真技术研究多阶段供应保障过程中的动态协调机制。以 Agent 之间的供需关系为基础给出了多 Agent 系统模型结构中小组的定义。为了给出备件短缺情况下使军事效益最大的备件分配策略，设计出以小组为单位的多 Agent 强化学习方法。最后通过仿真实例验证了方法的有效性。

关键词：战时备件供应保障系统；动态协调机制；多 Agent 系统
DOI：10.3969/j.issn.1009－3516.2009.03.013
中图分类号：TP391.9　　**文献标识码**：A　　**文章编号**：1009－3516(2009)03－0059－05

战时备件供应保障是为满足战场环境条件下航空装备对备件的需求，依托战时备件供应保障系统(Wartime Spares Support System，WSSS)所完成的备件储备供应过程中所涉及到的相关活动。战时备件供应保障的不确定性特点突出，包括备件需求过程的不确定和对需求满足过程的不确定。该特点使得精确保障要求下的战时备件供应保障不可能全面一步配置到位，需要根据作战任务、作战环境以及作战过程中备件供应保障系统的状态进行动态协调。类似的，备件供应保障专家 Alfredsson 明确地指出备件供应保障的目标是要实现基于状态的灵活备件供应保障[1]。

和平时期，采用仿真技术对信息化条件下的备件供应保障进行预先研究是认识战时备件供应保障的重要方法和手段。鉴于战时备件供应保障系统与多 Agent 系统的相似性，本文采用基于 Agent 的建模仿真技术研究多阶段供应保障过程中的动态协调机制，针对在作战实施过程中由于战场环境的动态性与需求的不确定性而出现不同需求节点之间有限备件资源的竞争，建立有效的动态协调机制，以发挥有限备件资源约束下的最大军事效益。

1　多 Agent 系统模型的建立

1.1　Agent 技术适用性分析

实际备件供应保障网络物理系统中节点具有分散特征且相对独立，随着计算机及信息网络能力的提高，不同节点具有一定的自主决策能力[2－3]。

多 Agent 系统(Multi Agent System，MAS)是由多个目标不同、行为方式不同的 Agent 组织在一起的，可以通过相互协作达到总体目标的 Agent 的集合。多 Agent 系统符合战时备件供应保障在实际运作过程中体现出来的自治性、分布性的特征。因此，可以采用多 Agent 系统有效地描述备件供应保障网络中节点之间的关

*　**收稿日期**：2009－01－15
作者简介：刘喜春(1979－)，女，内蒙古巴盟人，博士生，主要从事装备论证与仿真评估、备件供应保障研究
E－mail：olive－simple@sina.com

系和行为。表 1 给出了备件供应保障系统与多 Agent 系统的自然属性之间的相似性。

表 1 备件供应保障系统与多 Agent 系统的相似性

Tab. 1 The comparability between WSSS and MAS

WSSS 的自然属性	MAS 的自然属性
由不同层次的多个保障节点组成	由拥有不同角色和功能的不同类型的 Agent 组成
每个节点都有各自的目标、能力、所执行的特定任务,及所遵守的规则	Agent 有其自己的目标、资源、任务和他们的决策规则和程序
多阶段保障过程中需要根据实际作战情况在备件供应保障网络中协调备件	Agent 之间的交互是多 Agent 系统的基本行为模式
多阶段保障过程中备件供应保障结构在战场环境下可能发生变化	Agent 可以根据不同的控制和联接结构组织在一起,结构可变

1.2 多 Agent 系统模型的建立

建立多 Agent 系统模型时,要将 Agent 作为系统的基本抽象单元,赋予 Agent 一定的智能,然后在多 Agent 之间设置具体的交互方式,最终得到相应的系统模型。因此建立多 Agent 系统模型包括 2 个方面:一是单个 Agent 模型的建立,另一个是 Agent 之间的交互。

Agent 之间的交互以战时备件供应保障系统的供需关系为基础。按照 AGR 模型[4] 给出的 Agent、小组(Group)以及角色(Role)的概念,Agent 是一个可以在多个小组中扮演不同角色的自主个体。在备件供应保障结构中,节点之间的关系主要是供需关系,因此多 Agent 系统模型中 Agent 的角色可以是备件的"供应方",也可以是备件的"需求方",或二者的结合。用 $t = 1, 2, \cdots, T$ 表示有效的作战阶段,用二元组 $M(t) = (A(t), R(t))$ 表示多 Agent 系统在阶段 t 的供需关系,其中 $A(t)$ 表示所有 Agent 的集合,$R(t)$ 表示 Agent 之间的供需关系。对于 $A_i, A_j \in A(t)$,有下面几种关系:

$A_i \times A_j$,即 A_i 与 A_j 互为供应关系,二者形成供应回路;

$A_i \mathrm{O} A_j$,A_i 与 A_j 无关;

$A_i \wedge A_j$,A_i 供应 A_j;

$A_i \vee A_j$,A_j 供应 A_i;

A_i 的保障集为 $S_{\mathrm{set}_i}(t)$,指 A_i 在阶段 t 所保障的节点集合,有:$S_{\mathrm{set}_i}(t) = \{A_j \mid A_i \wedge A_j\}$。

按照 Agent 节点之间的供需关系可以把多 Agent 系统分为若干个小组,每个小组以扮演"供应方"角色的 Agent 为中心,为小组中其它"需求方"成员 Agent 提供备件供应。

参考王进发、李力等人对军事供应链的研究[5],柔性的战时备件供应保障系统结构是应对保障过程中不确定性的有效途径。多 Agent 系统模型的组织结构具有柔性特点,Agent 根据可能的供需关系组织在一起,能够很好地描述动态战场环境下系统结构的变化。例如战时某个备件供应保障节点毁伤后,对应的 Agent 节点退出多 Agent 系统模型,依据 Agent 之间的供需关系重新组成新的供应保障系统。

2 多 Agent 强化学习

2.1 多 Agent 协调机制

多 Agent 系统模型中通常采用的协调机制主要有 3 种:基于常规(或社会法则)的协调,基于通讯的协调以及基于学习的协调[6]。

强化学习(Reinforcement Learning, RL)[7] 作为一种无监督的学习方法,它考虑的是在没有外界指导的情况下,Agent 通过与不确定环境的交互从而获得最优解。对战时备件供应保障系统来说,环境就是瞬息万变的系统状态,强化学习的本质就是学习如何根据这些状态选择行为。图 1 给出了强化学习过程框图,强化学习基本过程包括:

图 1 强化学习过程框图

Fig. 1 The process of RL

1) Agent 通过通信模块获得环境的状态信息;

2) Agent 以某个决策规则选择一个动作或行动方案 a;

3) 下一时刻 Agent 从环境中获取一个奖赏值 r,以该奖赏值修正其内部的决策规则。

2.2　多 Agent 强化学习分类

多 Agent 系统中的强化学习可以根据同一时刻进行学习的 Agent 的数量分为主导 Agent 学习与多 Agent 共同学习[8]。

在主导 Agent 学习(又称为集中式强化学习)模型中,Agent 系统中每个时刻只有一个 Agent 在学习[9]。主导 Agent 以整个多 Agent 系统的整体状态作为输入,以对各个 Agent 的动作指派为输出,采用标准的强化学习方法,逐渐形成一个最优的协作机制。

多 Agent 共同学习模型中,同时有多个 Agent 共同学习,每个 Agent 都是学习的主体,强化学习过程是分布式的。每个 Agent 分别学习对环境的响应策略和相互之间的协作策略。

完全集中或完全独立的多 Agent 强化学习要么会造成状态空间和行动空间的组合爆炸,导致学习速度慢,要么不易达到全局最优[10]。本文给出基于分组强化学习的协调机制,Agent 之间协调在小组内部进行,小组外的 Agent 不直接参与,能够大大提高多 Agent 系统的协调效率。

3　分组强化学习

3.1　分组强化学习框架

与多 Agent 系统模型的柔性组织结构相对应,以小组为单位进行强化学习。小组中"供应方"Agent 作为主导 Agent,进行强化学习;"需求方"Agent 为一般 Agent,接受主导 Agent 的动作指派。图 2 给出了分组强化学习框架。

主导 Agent 强化学习的本质是通过学习获得如何在动态的战场环境中根据一般 Agent 传来的订购信息,结合自己的状态信息给出备件短缺情况下的备件分配策略。小组中,主导 Agent 与一般 Agent 并不是简单的控制与被控制的关系。一般 Agent 的主动性体现在通过奖赏值影响主导 Agent 在强化学习过程中行为的选择。

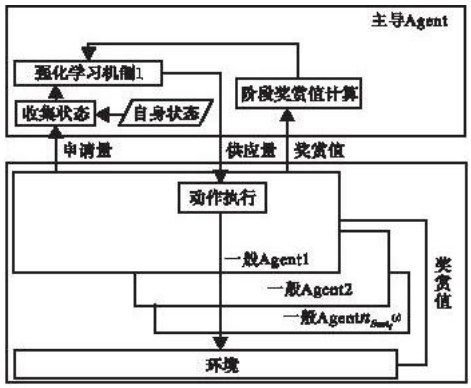

图 2　分组强化学习框架

Fig. 2　The structure of the RL by group

3.2　强化学习算法

主导 Agent 为保障集中每个一般 Agent 赋予权值,主导 Agent 根据权值分配备件。权值的大小表示相应的备件申请节点在竞争有限的备件资源时所体现出来的能力,权值大表明获得备件资源的能力强。图 3 给出了主导 Agent 强化学习模型框架。

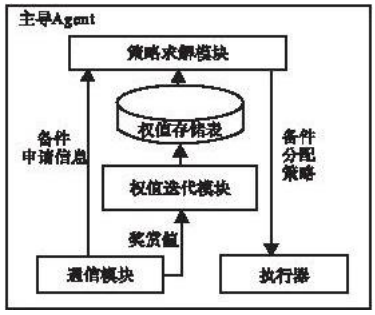

图 3　主导 Agent 强化学习模型框架

Fig. 3　The RL structure of the leading agent

在强化学习过程中,主导 Agent 获得一般 Agent 的奖赏值信息,根据权值迭代模块中的修正算法更新权值,并存储于权值存储表中。策略求解模块根据备件申请信息结合权值存储表中的权值信息给出备件分配策略。

多 Agent 系统模型中供应节点在备件库存不能满足所有备件申请时,采用分组强化学习给出备件分配策略,其目的是在有限的备件资源约束下发挥最大军事效益。本文以备件作为主要因素,不考虑其它因素对作战任务成功性的影响,因此采用表示备件供应保障水平的阶段备件满足率描述系统效能,作为强化学习的奖赏值。

4　仿真实例

为了验证多 Agent 强化学习的效果,在多 Agent 建模仿真平台 REPAST[11] 上进行仿真试验。强化学习实例中选取图 4(a)所示的 3 级备件供应保障结构,与实际 3 级备件供应保

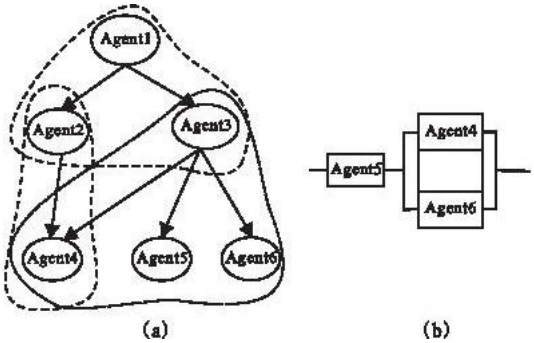

图 4　强化学习实例

Fig. 4　The example of RL

障结构对应,其中 Agent1 代表基地级库存节点,Agent2、Agent3 分别代表 2 个不同的中继级库存节点,Agent4、Agent5、Agent6 表示 3 个基层级作战单元。图 4 中 3 个虚线框表示 3 个不同的学习小组,例如以 Agent1 为主导 Agent 的小组中包括一般成员 Agent2 和 Agent3。

选取以 Agent3 为主导 Agent 的学习小组,用 $p_i(t)$ 表示 A_i 的阶段备件满足率,并用其来描述奖赏值。假设作战单元之间阶段作战任务成功概率的逻辑关系如图 4(b)所示,那么主导 Agent 的奖赏值为:

$$p_3(t) = p_5(t)(1-(1-p_4(t))(1-p_6(t)))$$

仿真过程中用自相关过程 AR(1)描述基层级节点 Agent4、Agent5、Agent6 的阶段备件需求数量为:

$$d(t) = d + \rho d(t-1) + \varepsilon(t)$$

取 $\rho = 0.8$,表 2 给出了各节点相关参数的取值。假设 Agent3 的阶段投入量服从均匀分布 $U(250, 300)$。

表 2 初始化表

Tab. 2 The initialization table

	初始需求	d	σ
节点 4	50	10	5
节点 5	100	20	10
节点 6	150	30	15

为了验证算法的有效性,将强化学习效果与平均随机分配策略进行比较。平均随机分配策略以需求节点申请备件数量所占总申请备件数量的比例为选择概率,按照轮盘赌的方式分配备件。记每个一般 Agent 的选择概率值 $p_m(t) = \dfrac{K_j(t)}{\sum_{l=1}^{n} K_l(t)}$,$K_j(t)$ 表示第 j 个 Agent 向主导 Agent 申请的备件数量。对于主导 Agent 中的每一个备件,先生成[0,1]内的随机数 r,若 $p_1^{(t)} + p_2^{(t)} + \cdots + p_{l-1}^{(t)} < r < p_1^{(t)} + p_2^{(t)} + \cdots + p_{l-1}^{(t)} + p_l^{(t)}$,则选择把该备件分配给 A_l。

图 5 中的曲线分别表示采用强化学习策略、平均随机分配策略时 Agent3 的奖赏值变化曲线。经过初期振荡后,强化学习策略明显优于平均随机分配策略。

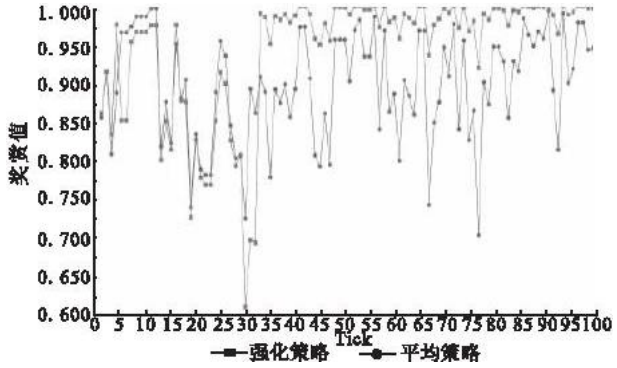

图 5 不同策略下的奖赏值变化曲线

Fig. 5 The plot of the reward under different strategy

5 结束语

战时备件供应保障不确定性特点突出,要对不确定环境进行快速响应,需要备件供应保障结构中的组成单元具有自治能力以及彼此协调能力[3]。本文采用基于 Agent 的建模仿真技术研究战时备件供应保障的动态协调机制,并建立了有效的分组强化学习方法,来解决多阶段作战过程中有限备件资源在系统中的协调问题。

本文通过仿真对精确保障要求下的战时备件供应保障动态协调机制进行预先研究。下一步研究以技术发展为基础,依托基于军队内联网的后勤 C^4I 系统,建立基于 Agent 的战时实时决策系统。此外,定性决策往往在战时处理某些偶然不确定事件方面更胜一筹,因此下一步研究可以在多 Agent 系统模型中把体现不同作战规则的专家知识作为重要知识以规则的形式添加到 Agent 的规则库中,用于指导协调策略的生成。

参考文献:

[1] Alfredsson P. Flexible Supply: The Next Step in the Evolution of Sparing Strategies[C]//SOLE 2000 35th Annual Proceedings,[S.l]:SOLE,2000.

[2] Lawson E, Ferris T, Cropley D,et al. Development of A Foundation for Military Network Science[R/OL]. [2009-4-2]. http://arrow.unisa.edu.au:8081/1959.8/47987.

[3] Kshanti Greene, David Cooper G, Michael Czajkowski, et al. A Cognitive Agent Architecture Optimized for Adaptivity [C]// DAMAS LNAI3890. Heidelberg:Spring Berlin,2006:104-120.

[4] Gutknecht J O, Michel F. From Agents to Organizations: An Organizational View of Multi-agent Systems[C]//AOSE Australia:AasE Melbourne,2003: 214-230.

[5] 王进发, 李励, 李仕明. 军事供应链的柔化结构[J]. 军事运筹与系统工程,2005, 19 (1):23-28.

WANG Jingfa, LI Li, LI Shiming. The Flexible Structure of Military Supply Chain[J]. Military Operations Research and System Engineering,2005, 19(1):23 - 28. (in Chinese)

[6] 郑淑丽, 韩江洪, 骆祥峰,等. 基于强化学习的多 Agent 协作研究[J]. 小型微型计算机系统, 2003, 24(11):1986 - 1988.
ZHENG Shuli, HAN Jianghong, LUO Xiangfeng, et al. Cooperative Multi - agent Systems Based on Reinforcement Learning [J]. Mini - Micro Computer Systems,2003,24(11):1986 - 1988. (in Chinese)

[7] Sutton R S, Barto A G.. Reinforcement Learning[M]. MA:MIT Press,1997.

[8] 仲　宇, 顾国昌, 张汝波. 多智能体系统中的分布式强化学习研究现状[J]. 控制理论与应用, 2003, 20(3):317 - 322.
ZHONG Yu, GU Guochang, ZHANG Rubo. Survey of Distributed Reinforcement Learning Algorithms in Multi - agent Systems [J]. Control Theory & Applications,2003,20(3):317 - 322. (in Chinese)

[9] Tan Ming. Multi - agent Reinforcement Learning: Independent vs Cooperative Agent[C]// In Proceedings of the 10th International Conference on Machine Learning (ICML - 93), San Fransisco:Morgan Kaufmann Publisher Inc,1993:487 - 494.

[10] 蔡庆生, 张　波. 一种基于 Agent 团队的强化学习模型与应用研究[J]. 计算机研究与进展, 2000, 37(9):1087 - 1093.
CAI Qingsheng, ZHANG Bo. An Agent Team Based Reinforcement Learning Model and its Application[J]. J of Computer Research and Development,2000, 37(9):1087 - 1093. (in Chinese)

[11] Tobias R, Hofmann C. Evaluation of Free Java - libraries for Social - scientific Agent Based Simulation[J/OL]. Journal of Artificial Societies and Social Simulation, 2004,7(1):[2009 - 4 - 1]. http://jasss. soc. surry. ac. uk/7/1/6. html.

(编辑:姚树峰,徐敏)

Research on Multi Agent Reinforcement Learning Based Dynamic Coordination Mechanism for Wartime Spares Support

LIU Xi - chun, WANG Chao, WANG Wen - guang, WANG Wei - ping
(Institute of Systems Engineering, School 5, National University of Defense Technology, Changsha 410073, China)

Abstract: Spare parts support plays an import role during wartime. In order to meet the requirements of Precision Support, spare parts support must be planned deliberatively prewar and be executed flexibly to deal with various uncertainties. Based on the similarity between the wartime spares support system and the multi agent system, Agent based modeling and simulation methods are adopted to investigate the dynamic coordinate mechanism during the wartime. Groups in the multi - agent system's structure are described on the bases of the relationship between the Agents. To decide how to supply spares dynamically during the wartime, the new multi agent reinforcement learning method is designed and presented. A simulation example is illustrated in the end and the simulation result shows that the method is effective.

Key words: wartime spares support system; dynamic coordination mechanism; multi agent system; reinforcement learning