文章编号: 1001-0920(2001)04-0465-04

遗传算法在逃逸机动策略中的应用研究

周 锐, 陈宗基
(北京航空航天大学 自动控制系, 北京 100083)

摘 要: 分析了基于强化学习原理和遗传算法的序贯决策规则的自动学习方法, 从规划报偿和规则激活度的角度讨论和研究了规则的信度分配问题, 解决了在大的状态空间中搜索和延迟评价问题, 为处理复杂的决策过程提供了一种行之有效的方法。基于该方法实现了飞机的逃避机动策略, 仿真结果表明了该方法的有效性。

关键词: 遗传算法; 强化学习; 逃逸机动策略
中图分类号: TP 18 **文献标识码**: A

Learning Evasive Maneuvers Using Genetic Algorithms

ZHOU Rui, CHEN Zong-ji
(Department of Automatic Control, Beijing University of Aeronautics and Astronautics, Beijing 100083)

Abstract: The method of learning sequential decision rules automatically based on reinforcement learning and genetic algorithms is analyzed. The credit assignment of rules is discussed from the point of view of plan payoff and rule strength. Based on the method, the problem of searching in a large state space and delayed critics is solved. The research provides a new approach for dealing with complex sequential decision making problems. The evasive maneuver games for plane in combat are implemented using the genetic algorithms and reinforcement learning. Simulation results show the effectiveness of the method.

Key words: genetic algorithms; reinforcement learning; evasive maneuver games

1 引 言

早期的逃逸策略主要基于微分对策理论[1], 但微分对策理论距实际应用还有一定的距离, 这就要求将新的理论引入微分对策, 以打破目前的僵局。一些学者将专家系统引入微分对策, 在一定程度上解决了微分对策理论的应用问题。目前, 知识的获取已成为专家系统设计的一个瓶颈, 而机器学习(例如强化学习[2]) 可使知识的获取过程自动化, 并扩展所能得到的知识库资源范围。这种过程自动化可通过强化学习的手段, 从仿真模型中学习知识或规则, 然后应用于实际的目标环境[3]。

学习过程实际上是一个在大的状态空间中搜索和优化的过程。遗传算法(GA) 由于具有鲁棒性和隐含并行性, 并能获得全局最优等特点[4], 为这种决策规则的学习和优化提供了一种强有力的手段。

收稿日期: 2000-04-29; 修回日期: 2000-11-07
基金项目: 国家自然科学基金项目(6990002); 航空基金项目(98D541102)
作者简介: 周锐(1968—), 男, 湖北钟祥人, 副教授, 从事飞行控制、制导与智能控制等研究; 陈宗基(1943—), 男, 上海人, 教授, 博士生导师, 主要从事鲁棒与自适应控制、人工智能与专家系统等研究。

2 规划自动学习原理

对于一个序贯决策过程, 从某一初始状态到达给定的状态或目标, 需要一序列控制作用或一组规则。将这一组控制作用序列称为一个策略(Policy), 与这组控制作用序列对应的一组规则称为一个规划(Plan)。显然, 一个规划确定了一个策略, 也就确定了从某一初始状态到达给定状态或目标的到达方式和代价。一个决策过程可有多种规划, 问题是如何自动确定最优规划, 或者说如何实现这些知识规则的自动学习和获取。

一种基于强化学习原理和遗传算法的规划自动学习原理如图 1 所示。它主要包括 3 个功能模块: 模块 A 主要实现世界建模的一些接口; 模块 B 主要通过读传感器、设定控制变量以及从评价中得到报偿等实现与世界模型的接口, 并执行规则的匹配和冲突的化解, 以及规则级的信度分配等; 模块 C 主要通过遗传算法的交叉和变异等操作, 从一些具有较高性能的规划中有选择地进行复制和修改, 以获得最优的规划及规则库。

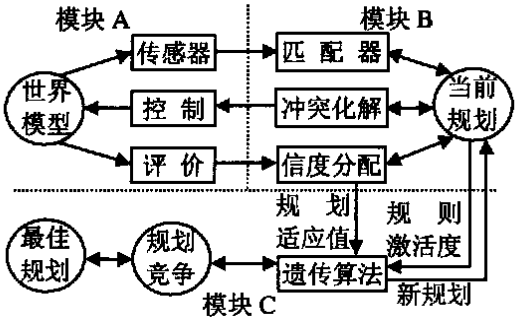

图 1 规划自动学习过程及原理

3 逃逸机动策略求解原理

3.1 逃逸对策

考虑导弹和飞机的二维追逃问题, 如图 2 所示。其中, M 和 T 分别表示导弹和飞机; L 为导弹的前提角, 前提角是导弹和飞机交汇三角形的理论角度; E_H 是前提角误差, 表示导弹相对于交汇三角形的偏差; n_T 和 n_c 分别是飞机的机动加速度和导弹的指令加速度。这里主要研究飞机的逃逸对策, 并假设飞机拥有足够的传感器以探测控制所必需的信息, 且信息范围做适当限制和量化。假设导弹按比例导引规律进行机动, 即

$$n_c = N' V_c \dot{\lambda} \tag{1}$$

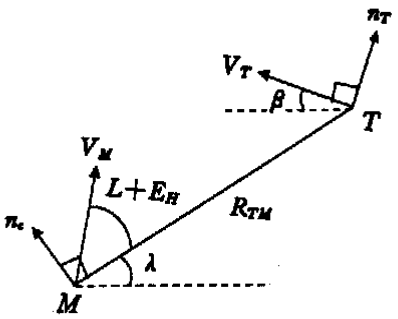

图 2 导弹和飞机二维空间关系

其中, N' 为导航比, V_c 为导弹与飞机之间的接近速度。飞机根据探测或估计到的系统主要状态信息, 这里主要选择 $V_M, V_T, E_H, R_{TM}, \lambda, n_T$, 由合适的规则确定自己的逃逸机动策略。规则一般形式为

$$\begin{aligned}&\text{IF } V_M^{\text{low}} \le V_M \le V_M^{\text{high}}, \quad V_T^{\text{low}} \le V_T \le V_T^{\text{high}}\\ &\quad R_{TM}^{\text{low}} \le R_{TM} \le R_{TM}^{\text{high}}, \quad n_T^{\text{low}} \le n_T \le n_T^{\text{high}}\\ &\quad E_H^{\text{low}} \le E_H \le E_H^{\text{high}}, \quad \lambda^{\text{low}} \le \lambda \le \lambda^{\text{high}}\\ &\text{THEN } n_T = n_T^{\text{opt}}\end{aligned}$$

一个对策过程包含很多这样的规则, 这些规则的集合称为一个规划。这些规则的上下限划分在开始时是随意的, 需要进行调整和优化。遗传算法的任务就是确定每条规则中状态变量最佳的上下限, 以及与每条规则相对应的最优机动策略。每条规则的染色体编码为

$$\begin{aligned}&[V_M^{\text{low}}\ V_M^{\text{high}}\ V_T^{\text{low}}\ V_T^{\text{high}}\ R_{TM}^{\text{low}}\ R_{TM}^{\text{high}}\ n_T^{\text{low}}\\ &\ n_T^{\text{high}}\ E_H^{\text{low}}\ E_H^{\text{high}}\ \lambda^{\text{low}}\ \lambda^{\text{high}} \mid n_T^{\text{opt}}]\end{aligned}$$

对于该问题的研究, 难点在于以下两方面:

1) 搜索空间: 虽然每个状态变量的量化是有限的, 但各状态组合后的搜索空间则是巨大的;

2) 延迟评价: 所有策略或规则的选取只有对策结束后才能评价其好坏, 如何评价中间过程对结果的影响程度, 则存在信度分配(Credit Assignment)问题。

采用遗传算法可以解决竞争学习问题, 而强化学习主要解决信度分配问题。信度分配主要在规则级, 遗传竞争主要在规划级。

3.2 信度分配

将一次对策过程称为一个仿真回合。它是指从飞机探测到导弹开始到导弹击中飞机, 或在规定的时间内导弹没有击中飞机(此时飞机逃逸成功) 为止这段仿真过程。在每次仿真回合结束时, 返回一个报偿或者利润(Payoff) 函数

$$r = \begin{cases} 1, & \text{飞机逃逸成功} \\ 0, & \text{飞机在某时刻被击中} \end{cases} \tag{2}$$

导弹可以在规定时间内任意时刻击中飞机。强化学习正是基于这种奖惩机制的学习方法。

为了研究信度分配问题,给每一条规则分配一个性能测度,称为规则激活度(Strength)。首先根据利润分享规划原理[5],在每个仿真回合结束时,根据评价单元提供的利润信号 r 估计与每条规则相关联的利润均值和方差。对于激活的规则 R_i,均值和方差更新规则为

$$\begin{cases} \mu_i(t) = (1-c)\mu_i(t-1) + cr \\ \sigma(t) = (1-c)\sigma_i(t-1) + c(\mu_i(t) - r)^2 \end{cases} \tag{3}$$

其中 $0 < c < 1$ 为利润分享速率。规则 R_i 的激活度定义为

$$\text{strength}_i(t) = \mu_i(t) - \sigma_i(t) \tag{4}$$

3.3　遗传算法学习原理

遗传算法应用于逃逸对策问题的学习原理及结构如图 3 所示[2]。初始群体由一定数量的初始规划组成,每个规划由一定数量的规则组成。每个规则初始化成最大的通用性,以使所有初始规则以相等概率匹配所有状态,并以均匀的概率被选中。整个系统主要包括推理系统和学习系统两部分。

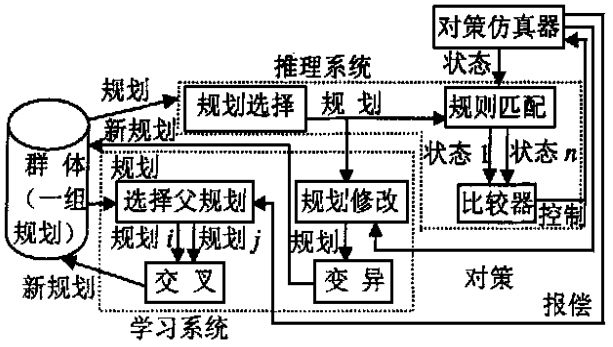

图 3　遗传算法学习原理及结构

3.3.1　推理系统

推理系统主要实现规则的匹配和规则的专一化等功能。规则匹配是对当前规划中规则集合进行考察,其中具有最佳匹配的规则被激活,并由该规则产生相应的控制量作用于对策系统。由于规则被初始化成匹配所有状态,因此在学习过程中需要对规则进行调整以适应特定的状态,即实现规则的专一化功能。规则的调整是根据状态变量对被激活规则的上下限进行修改,主要基于爬山法,即

$$\begin{cases} X_i^{\text{low}} = X_i^{\text{low}} + \alpha(X_i - X_i^{\text{low}}) \\ X_i^{\text{high}} = X_i^{\text{high}} + \alpha(X_i^{\text{high}} - X_i) \end{cases} \tag{5}$$

其中,α 为学习速率,X_i 分别代表状态变量 V_M, V_T, E_H, R_{TM}, λ, n_T。

3.3.2　学习系统

学习系统主要实现一个规划中规则的变异以及规划之间交叉等操作,它是根据适应值比例选择原则来选择进行变异的规则或进行交叉的规划。选中规则 R_i 进行变异的概率主要由该规则的激活度决定

$$P(R_i) = \frac{\text{strength}_i(t)}{\sum_{j \in \text{rules}} \text{strength}_j(t)} \tag{6}$$

某一规划被选中进行交叉的概率由该规划所产生的平均利润值确定,并用某一规划 plan_i 在一系列仿真回合中(假设 N 次试验)所产生的平均利润值来表示该规划的适应值,即

$$\text{fitness}(\text{plan}_i) = \frac{1}{N}\sum_{k=1}^{N} \text{payoff}(k) \tag{7}$$

则规划 plan_i 被选择进行交叉的概率为

$$P(\text{plan}_i) = \frac{\text{fitness}(\text{plan}_i)}{\sum_{j \in \text{plans}} \text{fitness}(\text{plan}_j)} \tag{8}$$

被选中的两个规划 plan_i 和 plan_j,其各自内部规则根据规则的激活度进行排序,从规划 plan_i 中选择一定数量的规则,再从 plan_j 中选择剩余数量的规则进行组合,形成新的规划取代群体中具有最小适应值的规划。

4　仿真研究

假设初始群体含 40 个规划,每个规划有 20 条规则。变异概率选择 0.01,交叉概率选择 0.8。共进行 10 000 次对策或回合的仿真试验,飞机逃逸成功的概率与仿真对策次数之间的关系曲线如图 4 所示。

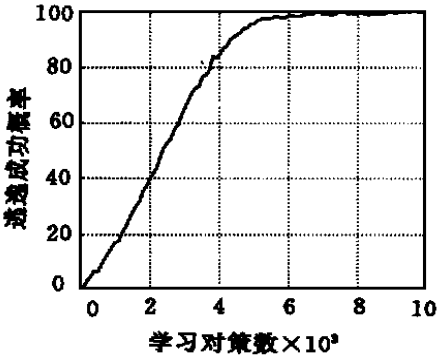

图 4　成功概率与试验次数之间的关系

从图中可以看出,试验次数达到 5 000 时,逃避成功的概率已达 95% 以上,达到了自动学习和获取规则的目的,表明了遗传算法在该问题中的有效性。

(下转第 472 页)

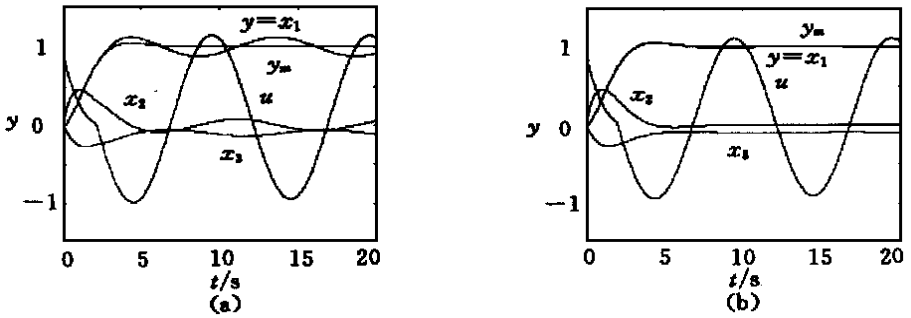

图 4 抗干扰能力

(a) 正弦干扰(L = 0.1s) (b) 正弦干扰(L = 0.01s)

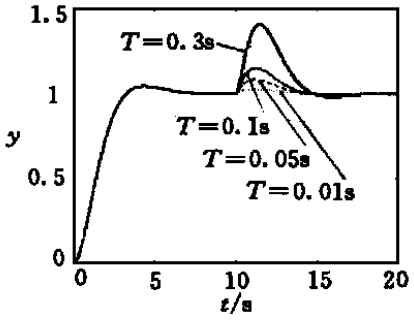

图 5 滤波器对系统性能的影响(L = 0.01s)

时的响应曲线。滤波器时间常数越小, 各项性能越好。

参考文献:

[1] 钟庆昌. 时滞控制及其应用研究[D]. 上海交通大学, 1999, 5(1): 45–49.

[2] 钟庆昌, 谢剑英, 梁春燕. 时滞滤波器及其应用研究[J]. 上海交通大学学报, 1999, 5(1): 45–49.

[3] Youcef Toumi K, Ito O. A time delay controller for systems with unknown dynamics [J]. J of Dynamic Systems Measurement & Control-Trans of the ASME, 1990, 112(4): 133–142.

[4] Chang Pyung H, Park Byung S, Park Ki C. Experimental study on improving hybrid position/force control of a robot using time delay control [J]. Mechatronics, 1996, 6(8): 915–931.

[5] Cheng Chi-Cheng, Chen Cheng-Yi. Controller design for an overhead crane system with uncertainty [J]. Control Engineering Practice, 1996, 4(5): 645-653.

[6] Park J H, Kim Y M, Yim J G. Time-delay sliding mode control for a servo system [A]. IEEE/ASME Int Conf on Advanced Intelligent Mechatronics[C]. AIM, 1997.

(上接 467 页)

本文仅仅讨论了飞机的单向对策问题, 其中还有不少需要进一步研究之处。如双向对策、规则的合并、规则的分解、规则的删除和添加等。此外, 关于报偿函数的选取和信度分配问题也是一个重要的研究课题。

参考文献:

[1] 张嗣瀛. 关于定量与定性微分对策[J]. 自动化学报, 1980, 6(2): 121–130.

[2] J W Sheppard. Multi-agent reinforcement learning[M]. Pittsburgh: Johns Hopkins University, 1997. 91–110.

[3] J Grefenstette, C Ramsey, A Schultz. Learning sequential decision rules using simulation models and competition [J]. Machine Learning, 1990, 5(4): 355–381.

[4] D E Goldberg. Genetic algorithms in search, optimization and machine learning[M]. Addison: Mass, 1989.

[5] J Grefenstette. Credit assignment in rule discovery systems based on genetic algorithms [J]. Machine Learning, 1988, 3(2/3): 225–245.