

DOI: 10.7652/xjtuxb201802010

采用双层强化学习的干扰决策算法

颢孙少帅^{1,2}, 杨俊安^{1,2}, 刘辉^{1,2}, 黄科举^{1,2}

(1. 国防科技大学电子对抗学院, 230037, 合肥; 2. 安徽省电子制约技术重点实验室, 230037, 合肥)

摘要: 为解决强化学习算法在干扰决策过程中收敛速度慢的问题,提出了一种采用双层强化学习的干扰决策算法(DRLJD)。首先对等效通信参数进行建模,模型减少了待学习参数的个数,降低了搜索空间的维度;然后利用降维后的搜索空间指导干扰参数选择,避免随机选择导致干扰性能差的缺点;最后利用选择的干扰参数施加干扰,并根据环境反馈进一步降低搜索空间的维度,通过不断交互的方式加快算法的收敛速度。此外,将以往的干扰经验以先验信息的形式加入到系统的学习进程中,进一步缩短了系统的学习时间。针对构造的干扰问题实验表明,DRLJD 算法经过 200 次交互便学习到优异的干扰策略,小于现有算法所需 600 次交互,且先验信息的利用进一步降低了对交互次数的要求。以提出的新的奖赏标准作为奖赏依据,算法能够在未知通信协议情况下以牺牲交互时间为代价学习到最佳干扰策略。

关键词: 强化学习;双层强化学习;干扰决策;先验信息;奖赏标准

中图分类号: TN975 **文献标志码:** A **文章编号:** 0253-987X(2018)02-0063-07

An Algorithm for Jamming Decision Using Dual Reinforcement Learning

ZHUANSUN Shaoshuai, YANG Junan, LIU Hui, HUANG Keju

(1. School of Electronic Countermeasure, National University of Defense Technology, Hefei 230037, China;

2. Key Laboratory of Electronic Restriction, Anhui Province, Hefei 230037, China)

Abstract: A novel algorithm for jamming decision using dual reinforcement learning (DRLJD) is proposed to accelerate convergence rate of reinforcement learning algorithms in jamming decision. First, a model of equivalent communication parameter is constructed to reduce both the number of unlearned parameters and the dimension of the search space. Secondly, the search space with reduced dimension is used to choose jamming parameters and to avoid worse jamming performance caused by random selection. Finally, the selected parameters are used to take jamming action, and to reduce the dimension of search space from the environment feedback information. The convergence rate of the algorithm is accelerated by constant interaction. Moreover, previous jamming experiences are used as prior information to further shorten the learning time of the system and to accelerate the convergence rate. The proposed DRLJD algorithm is validated by taking experiments on some jamming problems. Simulation results show that the algorithm obtains the optimal or suboptimal jamming strategy with 200 interaction times which is less than 600 interaction times of existing algorithms, and the use of prior information further reduces the requirements for the number of interactions. When the new reward standard is used as a basis for reward the proposed algorithm could learn the optimal jamming strategy at the expense of

收稿日期: 2017-05-22。 作者简介: 颢孙少帅(1990—),男,博士生;杨俊安(通信作者),男,教授,博士生导师。 基金项目: 安徽省自然科学基金资助项目(1308085QF99,1408085MKL46)。

网络出版时间: 2017-12-05 网络出版地址: <http://kns.cnki.net/kcms/detail/61.1069.T.20171205.2120.006.html>

interaction times in the case that communication protocols are not known.

Keywords: reinforcement learning; dual reinforcement learning; jamming decision; prior information; reward standard

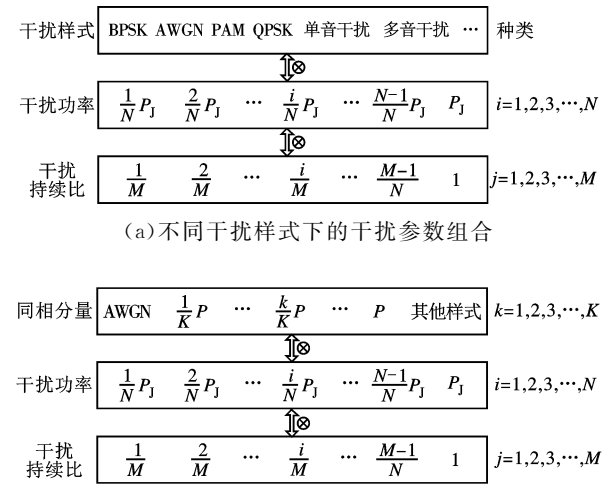
随着信息作战在现代化战争中的地位愈加巩固,确保己方夺取战场制信息权对取得局部战争乃至整场战役的胜利都是必不可少的。对于干扰敌方信息传递通道以实现通信拒止技术,其难度随着敌方在通信过程中使用认知无线电等技术而变得愈加困难。值得庆幸的是,无线通信的开放性使得对其干扰仍存在较大的可行性,而选择合适的干扰方式对实现成功干扰至关重要^[1-3]。许多参数优化理论算法如博弈论、粒子群算法、遗传算法等相继被用于搜索最优干扰策略。Wang 等人提出了一种通过观察可用频谱、信道质量、信道状态而采取的随机对策方法实现抗干扰通信^[4];Wu 等人分析了通信方在具备环境完全知识和部分知识的前提下,通过推导纳什均衡解来尽可能地降低干扰的影响^[5];Jin 等人构造了一种规则缩减遗传算法,通过对通信参数逐步优化的方式获得抗干扰策略^[6]。然而,上述理论算法成功实施的前提条件是需要具备通信方、环境的部分先验信息,无法适用于缺乏先验信息的场景。强化学习作为一种在线的、与外界环境实时交互的机器学习理论,具有噪声鲁棒性强、实时交互、无需先验信息的特点,普遍用于游戏^[7]、机器人控制^[8]等领域。将强化学习相关算法用于策略的学习之中,能够充分利用算法无需先验信息的优势,并且通过持续交互获得的学习结果接近外界真实环境状态。Gwon 等人提出利用 Q 学习的方法求取最优信道接入策略^[9];Slimeni 针对认知无线网络中的干扰问题,以改进的 Q 学习方法对抗固定的干扰策略,得出一种多信道功率分配方案^[10];Amuru 等人深入研究了利用强化学习算法从物理层、介质访问控制(media access control,MAC)层搜索最优干扰策略^[11-12];黄科举等人对 IEEE 802.11MAC 层认知干扰技术进行了研究,对干扰参数和干扰策略进行在线优化^[13]。但是,上述强化学习算法面临收敛速度慢的困境,在学习过程中需要同环境进行大量交互。虽然有一些相关算法^[8,14]专注于解决此类问题,但是尚没有具体算法研究强化学习在干扰决策过程中利用先验信息。

为了解决上述问题,本文提出了一种采用双层强化学习的干扰决策(jamming decision using dual reinforcement learning,DRLJD)算法。该算法在对

干扰问题建模后,利用环境反馈信息学习等效通信参数,然后利用等效通信参数学习干扰参数,通过两者交互学习的方式加快收敛速度。通过仿真结果可知,双层强化学习算法能够降低寻优过程中的交互次数,且先验信息的加入进一步加快了算法的收敛速度,提出的新的奖赏标准能够在未知通信协议的情况下以牺牲交互时间为代价学习到最佳干扰策略。

1 干扰模型

寻找最佳干扰参数问题往往被建模成多臂老虎机(multi-armed bandit,MAB)问题^[11]。以使用尽可能小的干扰功率实现期望符号错误率(symbol error rate,SER) ζ_E 为例,干扰信号与通信信号频率对准后,首先对干扰功率、信号调制样式、干扰持续比等干扰参数进行离散化处理,其次对各种可能的组合逐一尝试便可确定最佳组合,即最佳干扰策略,图 1 给出了不同干扰参数离散组合策略。图 1a 中以干扰样式、干扰功率、干扰持续比 3 种变量的组合构造干扰策略,其中干扰样式包括二进制相移键控(binary phase shift keying,BPSK)、加性高斯白噪声(additive white Gaussian noise,AWGN)、脉冲幅度调制(pulse amplitude modulation,PAM)、正交相移键控(quadrature phase shift keying,QPSK)等,干扰功率被划分成 N 个不同的功率等级,干扰持续比 ρ 为干扰时间与通信时间的比值,同样按设



(a)不同干扰样式下的干扰参数组合
(b)不同同相分量下的干扰参数组合
图 1 不同干扰参数离散组合策略

定值 M 对其进行均匀划分。图 1b 从同相分量、干扰功率、干扰持续比的角度构造不同的干扰策略。

直接估计干扰信号在接收端的等效功率是比较困难的,但可以由干扰方学习等效通信参数组合,并根据学习的通信参数组合来指导干扰参数组合的选择。对于该等效通信参数,有一个基本条件需要满足:即等效通信参数的取值是恒定或唯一的,不会因干扰功率的变化而变化。

假定通信信号调制样式为多电平正交幅度调制(multiple quadrature amplitude modulation, MQAM),同相维(I路)的维度为 X ,正交维(Q路)的维度为 Y 。通信信号发射功率 $P_T = \{P_T^{(1)}, P_T^{(2)}\}$,其中上标(1)和(2)分别表示参数在 I 路和 Q 路的分量,传输损耗因子为 α ,干扰信号发射功率 $P_J = \{P_J^{(1)}, P_J^{(2)}\}$,干扰持续比为 ρ ,传输损耗因子为 β ,信道噪声为均值为 0、方差为 1 的加性高斯白噪声,接收方使用天线自适应调零技术,使得干扰信号 I 路和 Q 路的衰减因子为 $\gamma = \{\gamma^{(1)}, \gamma^{(2)}\}$,判决并纠错后,错误率减少至实际值的 θ 倍, $0 < \theta \leq 1$ 。接收端符号错误率 ζ_R 可表示如下

$$\zeta_R = [1 - (1 - \zeta_R^{(1)})(1 - \zeta_R^{(2)})]\theta\rho \quad (1)$$

$$\zeta_R^{(1)} = \frac{(X-1)}{X} \left\{ Q\left[(\beta P_T^{(1)})^{1/2} + (\alpha\gamma^{(1)} P_J^{(1)})^{1/2}\right] + Q\left[(\beta P_T^{(1)})^{1/2} - (\alpha\gamma^{(1)} P_J^{(1)})^{1/2}\right] \right\} \quad (2)$$

$$\zeta_R^{(2)} = \frac{(Y-1)}{Y} \left\{ Q\left[(\beta P_T^{(2)})^{1/2} + (\alpha\gamma^{(2)} P_J^{(2)})^{1/2}\right] + Q\left[(\beta P_T^{(2)})^{1/2} - (\alpha\gamma^{(2)} P_J^{(2)})^{1/2}\right] \right\} \quad (3)$$

符号错误率 ζ_I 可以表示如下

$$\zeta_I = [1 - (1 - \zeta_I^{(1)})(1 - \zeta_I^{(2)})]\theta\rho \quad (4)$$

$$\zeta_I^{(1)} = \frac{(X-1)}{X} \left\{ Q\left[(\bar{P}_T^{(1)})^{1/2} + (P_J^{(1)})^{1/2}\right] + Q\left[(\bar{P}_T^{(1)})^{1/2} - (P_J^{(1)})^{1/2}\right] \right\} \quad (5)$$

$$\zeta_I^{(2)} = \frac{(Y-1)}{Y} \left\{ Q\left[(\bar{P}_T^{(2)})^{1/2} + (P_J^{(2)})^{1/2}\right] + Q\left[(\bar{P}_T^{(2)})^{1/2} - (P_J^{(2)})^{1/2}\right] \right\} \quad (6)$$

对于干扰方而言,等效通信功率假定为 $\bar{P}_T$,利用一定的侦察手段获取通信信号的调制样式,并令式(4)计算的符号错误率 ζ_I 等于环境真实反馈值 ζ_R ,并以此限定参数 $\bar{P}_T^{(1)}$ 、 $\bar{P}_T^{(2)}$ 的取值范围, $\bar{P}_T^{(1)}$ 和 $\bar{P}_T^{(2)}$ 表示 $\bar{P}_T$ 在 I 路和 Q 路的功率值。

在式(1)~(3)中,参数 P_T 、 α 、 β 、 γ 、 θ 、 $P_T^{(1)}$ 、 $P_T^{(2)}$ 为固定未知量, ζ_R 仅与变量 $P_J^{(1)}$ 、 $P_J^{(2)}$ 以及 ρ 相关,

换言之,式(4)中 ζ_I 同样仅与 $P_J^{(1)}$ 、 $P_J^{(2)}$ 、 ρ 相关,与 $P_T^{(1)}$ 、 $P_T^{(2)}$ 无关,而当 $P_T^{(1)}$ 、 $P_T^{(2)}$ 为恒定值时, $\bar{P}_T^{(1)}$ 、 $\bar{P}_T^{(2)}$ 同样为恒定值。

以学习等效通信参数的方式取代对干扰参数的学习,学习的参数由同相分量、干扰功率、干扰持续比变为同相分量、等效通信功率。对上述参数进行离散化处理,如将同相分量、干扰持续比划分成 51 个等级,将干扰功率、等效通信功率划分成 21 个等级,此时待学习的参数组合数量由 $51 \times 21 \times 51 = 54\,621$ 个降至 $51 \times 21 = 1\,071$ 个。由于等效通信参数为恒定值,故 1 071 个可选参数组合中必然有一个组合最接近 $\bar{P}_T^{(1)}$ 、 $\bar{P}_T^{(2)}$,利用交互学习的方式学习该参数组合,使得搜索空间的维数得到极大降低,有利于减少学习过程中所需的交互次数。

2 双层强化学习算法

2.1 双层强化学习模型

双层强化学习的核心思想是在传统强化学习系统中增加一个学习模块,即不仅对等效通信参数进行学习,同时还对干扰参数进行学习。双层强化学习模型如图 2 所示,主要包括 2 个学习模块和一个策略选择模块。相对于传统的强化学习方法,该模型不仅能够根据强化学习的自学习过程实现干扰参数模块和等效通信参数模块的双层学习,还能够对以往的干扰经验加以利用,加快学习模块的收敛。

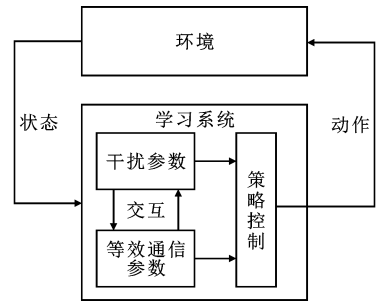


图2 双层强化学习模型

干扰参数模块 S_J 和通信参数模块 S_T 内包含参数离散化后的各种组合,本文选择的离散化处理方式与图 1b 类似,干扰参数模块内离散化处理对象是功率 P_J 、干扰持续比 ρ 、功率同相分量 $P_J^{(1)}$ 等参数,通信参数模块内离散化处理对象是功率 P_T 、功率同相分量 $P_T^{(1)}$ 等参数。

2.2 奖赏标准

本文对通信信号施加干扰时,以接收方获取信息的 SER 作为奖赏依据,该信息可通过对环境的侦察获得,例如美军战术互联网以 TCP/IP 协议作为

电台间的通信协议,该协议要求接收方在顺利接收数据包后发送确认帧/非确认帧(acknowledge/not acknowledge,ACK/NACK)信息,统计该信息后可用于计算包错误率(packet error rate,PER) R_{PE} ,而 R_{PE} 与 ζ_E 是等效的,可利用公式 $\zeta_E = 1 - (1 - R_{PE})^{1/H}$ 计算获得,其中 H 表示校验帧比特数。

在不具备有关通信信号相关先验信息时,可通过分析通信信号的行为评估干扰效果,本节提出一种新的奖赏标准——忍受时间。忍受时间定义为:通信方从受到干扰的时间为起始,到改变通信参数进行通信的时间为终止。显然,如果干扰参数性能优异,必然在短暂时间内迫使通信方改变通信参数,而性能较差的参数带来的包错误率相对较低,通信方能够忍受更长时间的干扰。将忍受时间作为奖赏标准,忍受的时间越久,奖赏值越低。此外,将其作为奖赏标准的好处是,无论通信方增大信号功率还是改变信号调制样式,其受干扰时忍受时间是确定的,该时间能够衡量干扰参数性能的优劣。

2.3 不同程度的先验信息

先验信息的加入有利于缩短学习过程所需的交互次数,为算法走向实用性提供保证。本文采用以下4种不同的先验信息:①精准的先验信息,干扰方能够在较低的干扰功率下达到期望的干扰效果,但该先验信息并非最优干扰策略;②含糊的先验信息,该信息唯一能够确定的是其满足干扰需求;③多份先验信息,以往的诸多干扰经验可作为多份先验信息;④错误的或无先验信息,该信息属于以往的成功干扰策略,但满足不了当前的干扰任务,达不到预期的干扰效果,可将该情况同样视为不存在先验信息。

2.4 采用双层强化学习的干扰决策算法

DRLJD算法可分为2个阶段,初始化阶段和主循环阶段。在初始化阶段,干扰方从干扰参数组合中随机选择一组 $\{P_J, P_J^{(1)}, \rho\}$,构造干扰信号对通信信号施加干扰,并优先利用先验信息构造干扰信号。根据反馈信息 ζ_R 对通信参数模块内的组合进行剔除操作,考虑到较大值的通信功率必然导致符号错误率更低,干扰方需要对 S_T 进行更新,将符号错误率低于 ζ_R 的通信参数组合删除。更新方式如下

$$S_T =$$

$$\{(\bar{P}_T^{(1)}, \bar{P}_T^{(2)}) \mid \zeta(P_J, P_J^{(1)}, \rho, \bar{P}_T^{(1)}, \bar{P}_T^{(2)}) < \zeta_R\} \quad (7)$$

式中: $\zeta(P_J, P_J^{(1)}, \rho, \bar{P}_T^{(1)}, \bar{P}_T^{(2)})$ 由干扰方利用给定参数并根据式(4)计算获得。在主循环阶段,干扰方根据等效通信参数模块中的可能解,首先从干扰参数

模块中根据式(8)选择合适的干扰参数组合 $\{P_J, P_J^{(1)}, \rho\}$ 生成干扰信号并实施干扰,然后采取与初始化阶段相同的操作步骤,对等效通信参数模块内的组合进行剔除操作。干扰参数选择方式如下

$$S_J = \{(P_J, P_J^{(1)}, \rho) \mid \zeta(P_J, P_J^{(1)}, \rho, \bar{P}_T^{(1)}, \bar{P}_T^{(2)}) > \zeta_E\} \quad (8)$$

式中: $\bar{P}_T^{(1)}, \bar{P}_T^{(2)}$ 取等效通信参数模块内功率最大的组合。最后,干扰方根据等效通信参数模块中剩余的可能组合和反馈信息 ζ_R ,对干扰参数模块内的组合进行剔除操作,删除不合理组合,如下式所示

$$S_J = \{(P_J, P_J^{(1)}, \rho) \mid \zeta(P_J, P_J^{(1)}, \rho, \bar{P}_T^{(1)}, \bar{P}_T^{(2)}) < \zeta_E\} \quad (9)$$

式中: $\zeta(P_J, P_J^{(1)}, \rho, \bar{P}_T^{(1)}, \bar{P}_T^{(2)})$ 由干扰方根据式(4)计算获得。DRLJD算法的详细步骤如下。

步骤1 对干扰信号、等效通信参数进行离散化处理,确定参数及变量的取值范围,给定初始化阶段以及主循环阶段的干扰次数 T_1 和 T ,明确先验信息。

步骤2 for $t=1:T_1$ 执行以下(1)~(2)步:

(1)从干扰参数组合中随机选择一组策略,构造干扰信号并对通信信号施加干扰,如果存在先验信息,则直接对其加以利用;

(2)根据反馈信息对等效通信参数模块内的组合进行剔除操作,删除不合理组合。

步骤3 for $t=T_1+1:T$ 执行以下(1)~(3)步:

(1)根据等效通信参数模块中的可能解,从干扰参数模块中选择合适的干扰参数组合;

(2)根据反馈信息对等效通信参数模块内的组合进行剔除操作,删除不合理组合;

(3)根据等效通信参数模块中的可能解和反馈信息,对干扰参数模块内的组合进行剔除操作,删除不合理组合。

2.5 算法性能指标

本文确定的干扰目标是,如何在满足期望 ζ_E 的情况下使得干扰功率尽可能低,且学习最优干扰参数所需的交互次数尽可能少。为有效衡量DRLJD算法的性能,本文主要从算法学习到最优(次优)干扰参数的概率、收敛所需干扰次数以及新的奖赏标准的可行性3个方面进行评估。

3 实验仿真

假定干扰对象为矩形QPSK信号,信号功率为 $P_T=100\text{ W}$ 、 $\alpha=0.9$ 、 $\theta=0.85$ 、 $\{\gamma^{(1)}, \gamma^{(2)}\}=\{0.35, 0.3\}$ 、 $P_J \in [100, 400]\text{ W}$ 、 $\bar{P}_T \in [100, 300]\text{ W}$,将 P_J 、

$\bar{P}_T$ 划分成 21 个等级, ρ 划分成 51 个等级, $\beta=0.8$, 对 $P_J^{(1)}$ 、 $\bar{P}_T^{(1)}$ 同样进行划分, 期望 $\zeta_E=0.38$ 。信道噪声为均值为 0、功率为 1 W 的加性高斯白噪声, 学习过程中以符号错误率作为奖赏依据。

3.1 学习结果

本文所提 DRLJD 算法包括初始化和主循环 2 个阶段, 本节以 20 次交互为增长步长, 就初始化交互次数介于 20 次和 200 次之间的情况展开讨论, 实验次数为 500 次。构造的先验信息 $\{P_J, P_J^{(1)}, \rho\}$ 如下: ①精准的先验信息, $\{180, 180, 1\}$; ②含糊的先验信息, $\{380, 190, 1\}$; ③多份的先验信息, $\{380, 180, 1\}, \{180, 180, 1\}, \{180, 0, 1\}, \{190, 190, 1\}, \{190, 0, 1\}$; ④错误的先验信息 $\{100, 50, 1\}$ 。根据 MAB 算法^[11]可知, 最优干扰功率为 160 W, 次优干扰功率为 170 W, 并且干扰功率主要集中在 I 路, 本节以 DRLJD 算法学习到最优和次优干扰策略的概率之和衡量算法性能。图 3 给出了先验信息及初始化交互次数对本文算法学习结果的影响。

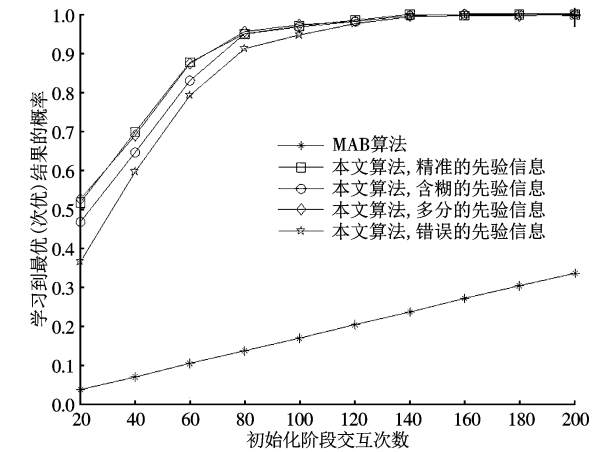


图 3 先验信息及初始化交互次数对学习结果的影响

从图 3 中可以看出, MAB 算法至少需要 600 次交互, 算法才能以概率 1 学习到最优和次优干扰策略。对比不同先验信息下的学习结果可以发现, 精准的先验信息和多份先验信息拥有相似的学习结果, 且概率较含糊的先验信息和错误的先验信息要大, 表明先验信息越精确, 学习结果越优异, 同时也说明学习结果与先验信息的个数无关, 仅与其准确程度息息相关。随着初始化阶段交互次数的增加, 4 类先验信息学习到最优和次优结果的概率趋于 1。

3.2 收敛速度

DRLJD 算法的收敛速度主要从主循环阶段进行分析, 表 1 给出了不同程度的先验信息及初始化交互次数对算法收敛次数的影响, 实验条件与 3.1

节相同, 收敛结果的准确性与图 3 相对应。

表 1 先验信息及初始化交互次数对本文算法收敛次数的影响

初始化交互次数	算法收敛次数			
	精准先验信息	含糊先验信息	多份先验信息	错误先验信息
20	23.55	24.92	25.93	25.51
40	43.91	44.45	44.96	44.34
60	63.73	63.96	64.15	64.13
80	83.48	83.39	83.45	83.73
100	103.86	103.53	103.35	103.61
120	123.44	123.24	123.38	123.36
140	143.40	143.21	143.43	143.16
160	163.38	163.33	163.30	163.48
180	183.13	183.18	183.15	183.13
200	203.14	203.20	203.20	203.06

从表 1 中横向比较可以看出, 无论干扰方具备何种程度的先验信息, 在相同的初始化交互次数时, 算法的收敛次数近乎是相等的。这是因为经过了初始化阶段的预学习, 等效通信模块和干扰模块内的可行解已经被限定在固定的范围内, 所以在主循环阶段, 算法仅需要额外的几次交互便收敛到固定解。从纵向比较来看, 在初始化阶段结束后, 本文算法平均需要额外的 3~6 次交互便收敛到稳定解, 反映出算法的主循环阶段学习能力不足, 仅能够剔除少量不合理解, 尚不具备进一步学习的能力。

结合图 3 与表 1 中的学习结果, 可以得出以下结论: DRLJD 算法学习到最优或次优干扰策略时所需的交互次数远远小于 MAB 算法所需的次数, 若想要进一步缩短交互次数, 便只能以牺牲学习结果的准确性为代价换取交互次数的进一步降低。在执行干扰任务时, 干扰方需要具备对敌方通信信号进行快速、准确干扰的能力。由于通信方在受到一定时间的干扰后, 会改变信号调制样式等操作来对抗干扰行为, 此时干扰方将面临新的学习任务, 需要重新开始学习, 因此干扰方需要在尽可能短的时间内, 实现对目标信号的顺利干扰, 即使存在干扰性能更加优异的策略, 同迫在眉睫的干扰任务相比, 可以适当牺牲学习结果的准确性。学习到次优的干扰策略仅仅意味着使用的干扰功率值较大, 主动权在干扰方, 属于可控因素, 但长时间的交互意味着需要通信状态保持不变, 主动权在通信方, 对干扰而言属于不可控因素。因此, 从实际情况出发, 牺牲学习结果

的准确性换取交互次数的减少是值得的,有利于尽可能早的对通信信号进行压制以实现通信拒止。

3.3 新的奖赏标准——忍受时间

本节实验条件同 3.1 节相一致,但仅对等效通信功率进行学习,并假定功率在 I 路和 Q 路上平均分配,奖赏标准以忍受时间替代符号错误率,持续干扰时间为 100 个数据包长度。等效通信信号功率取值范围介于 $[100\text{ W}, 300\text{ W}]$ 。

图 4 给出了离散化处理对算法学习性能的影响。由图 4a 可见,当离散值较小时,如小于 10,无法保证算法以概率 1 学习到最优或次优干扰策略,原因在于离散化处理过于粗糙,致使学习到的等效通信功率值不具有参考价值;图 4b 给出了离散化处理对算法收敛次数的影响,可以看出,划分越详细,所需要的收敛次数就越多,即学习时间越长;图 4c 给出了离散化处理对交互次数的影响,可以看出,交互次数等于离散后等效通信功率的个数。根据实验结果可以得出:忍受时间能够作为干扰策略学习过程中的奖赏标准,不足之处在于延长了算法的学习时间。

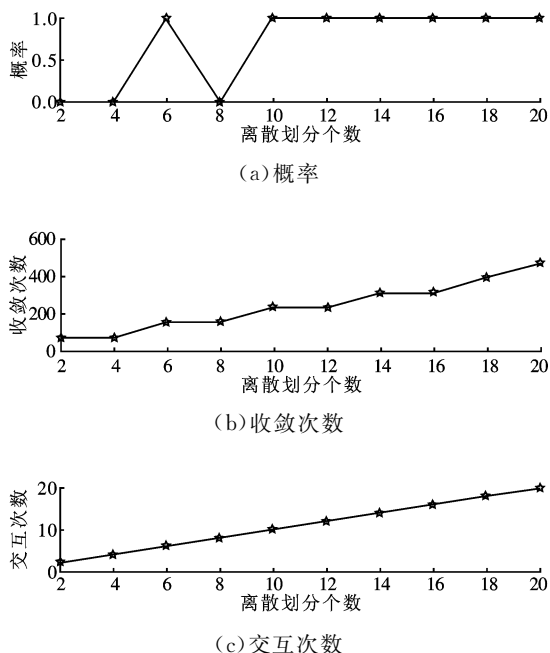


图 4 离散化处理对算法学习性能的影响

3.4 算法比较

为增加本文算法性能说服力,将本文所提 DRLJD 算法与 MAB 算法、Greedy 算法作比较,并从算法收敛所需交互次数、干扰性能两个方面对实验结果进行评价,图 5 给出了不同算法的收敛曲线。

从图 5 中可以看出,DRLJD 算法将初始化阶段

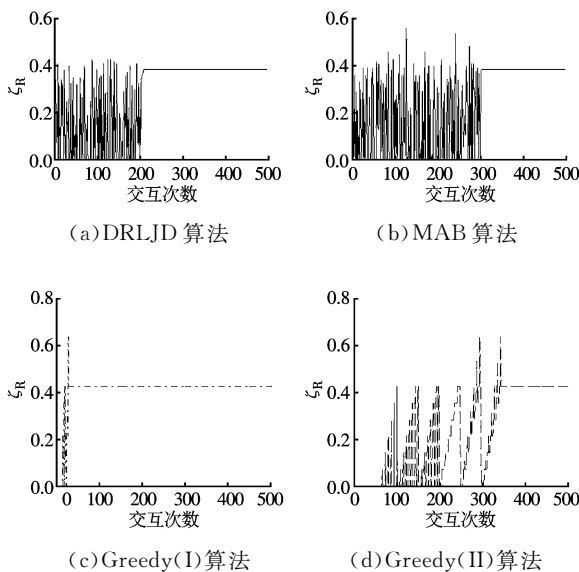


图 5 不同算法的收敛曲线

设置为 200 次交互,在主循环阶段,经过少量交互便收敛到 $\zeta_R = 0.3825$,接近任务期望值 $\zeta_E = 0.38$ 。MAB 算法需要将 54 621 种可能的干扰策略逐一尝试,才能确定最佳干扰策略,为便于比较,在实验过程中该算法首先进行 300 次交互,并从中选择效果最好的动作作为学习结果,在该条件下,算法收敛到 $\zeta_R = 0.3912$ 。Greedy(I)算法和 Greedy(II)算法将待学习的干扰参数分别划分成 3 个等级和 7 个等级,相应得到 27 个和 343 个干扰动作,需要逐一尝试后才能挑选出最佳动作,尽管划分不同,但两者均收敛至 $\zeta_R = 0.4254$,说明参数划分分辨率依然太低。因此,在干扰任务中,DRLJD 算法能够学习到比 MAB 算法、Greedy 算法更优的干扰策略,且需要更少的交互次数。

4 结 语

本文提出了一种采用双层强化学习的干扰决策算法,该算法在构造的双层强化学习模型基础上,同时对等效通信参数和干扰参数进行学习,通过预测的方式选择干扰策略,以降低学习效果为代价加快算法学习速度。提出的新的奖赏标准尽管在学习能力方面弱于原有标准,但因无需获悉网络协议使得算法适用领域更加广泛。此外,算法具有很强的扩展性,能够对已知先验信息加以利用,进一步缩短算法所需的交互次数。

参考文献:

- [1] 张春磊, 杨小牛. 认知电子战与认知电子战系统研究[J]. 中国电子科学研究院学报, 2014, 9(6): 551-

- 555.
- ZHANG Chunlei, YANG Xiaoniu. Research on the cognitive electronic warfare and cognitive electronic warfare system [J]. Journal of CAEIT, 2014, 9(6): 551-555.
- [2] 贾鑫, 朱卫纲, 曲卫, 等. 认知电子战概念及关键技术 [J]. 装备学院学报, 2015, 26(4): 96-100.
- JIA Xin, ZHU Weigang, QU Wei, et al. Concept of cognitive electronic warfare and it's key technology [J]. Journal of Equipment Academy, 2015, 26(4): 96-100.
- [3] YANG D, XUE G, ZHANG J, et al. Coping with a smart jammer in wireless networks: a Stackelberg game approach [J]. IEEE Transactions on Wireless Communications, 2013, 12(8): 4038-4047.
- [4] WANG Beibei, WU Yongle, LIU K J R. An anti-jamming stochastic game in cognitive radio networks [J]. IEEE Journal on Selected Areas in Communications, 2011, 29(4): 877-889.
- [5] WU Yongle, WANG Beibei, LIU K J R, et al. Anti-jamming games in multi-channel cognitive radio networks [J]. IEEE Journal on Selected Areas in Communications, 2012, 30(1): 4-15.
- [6] JIN H, SONG X Q, WANG M, et al. A fast anti-jamming decision method based on the rule reduced genetic algorithm [J]. KSII Transactions on Internet and Information Systems, 2016, 10(9): 4549-4567.
- [7] WANG H, HUANG T, LIAO X, et al. Reinforcement learning in energy trading game among smart micro-grids [J]. IEEE Transactions on Industrial Electronics, 2016, 63(8): 5109-5119.
- [8] XIA Wei, LI Huiyun, LI Baopu. A control strategy of autonomous vehicles based on deep reinforcement learning [C] // Proceedings of 2016 9th International Symposium on Computational Intelligence and Design. Piscataway, NJ, USA: IEEE, 2016: 16637586.
- [9] GWON Y L, DASTANGOO S, FOSSA C E, et al. Competing mobile network game: embracing antijamming and jamming strategies with reinforcement learning [C] // 2013 IEEE Conference on Communications and Networks Security. Piscataway, NJ, USA: IEEE, 2013: 28-36.
- [10] SLIMENI F, SCHEERS B, NIR V L, et al. Learning multi channel power allocation against smart jammer in cognitive radio networks [C] // 2016 International Conference on Military Communications and Information Systems. Piscataway, NJ, USA: IEEE, 2016: 7496544.
- [11] AMURU S, TEKIN C, SCHAAR M, et al. Jamming bandits: a novel learning method for optimal jamming [J]. IEEE Transactions on Wireless Communications, 2016, 15(4): 2792-2808.
- [12] AMURU S, BUEHRER R M. Optimal jamming using delayed learning [C] // Proceedings of 33rd Annual IEEE Military Communications Conference. Piscataway, NJ, USA: IEEE, 2014: 1528-1533.
- [13] 黄科举, 杨俊安, 周继航, 等. 基于多臂赌博机模型的 IEEE 802.11 MAC 协议认知干扰技术 [J]. 通信对抗, 2017, 36(2): 30-35.
- HUANG Keju, YANG Junan, ZHOU Jihang, et al. Cognitive jamming in IEEE 802.11 MAC based on multi-armed bandits [J]. Communication Countermeasures, 2017, 36(2): 30-35.
- [14] CHU C Y, ITO S, HARADA T, et al. Position-based reinforcement learning biased MCTS for general video game playing [C] // IEEE Conference on Computational Intelligence and Games. Piscataway, NJ, USA: IEEE, 2017: 7860449.

[本刊相关文献链接]

- 李清伟, 郭黎利. DS-CDMA 系统多波形优化的多址干扰抑制方法. 2017, 51(10): 94-99. [doi: 10.7652/xjtuxb201710016]
- 刘立, 张衡阳, 毛玉泉, 等. 变换域通信系统抗干扰编码幅度谱成型算法. 2017, 51(2): 91-96. [doi: 10.7652/xjtuxb201702015]
- 徐张宝, 马大为, 姚建勇, 等. 采用干扰估计的液压系统自适应鲁棒控制. 2016, 50(8): 123-129. [doi: 10.7652/xjtuxb201608020]
- 赵凯, 王闯, 李尊朝, 等. 结合平衡和滤波技术抑制 GaN 电源转换器的电磁干扰. 2016, 50(2): 38-42. [doi: 10.7652/xjtuxb201602007]
- 孙黎, 徐洪斌. 协作式终端直通系统中星座旋转辅助的干扰避免策略. 2015, 49(12): 6-11. [doi: 10.7652/xjtuxb201512002]
- 郭一鸣, 彭华, 张冬玲, 等. 联合串行干扰抵消与因子图的单通道混合信号盲分离算法. 2015, 49(10): 130-135. [doi: 10.7652/xjtuxb201510021]

(编辑 刘杨)