(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 115619607 A
(43)申请公布日 2023.01.17

(21)申请号 202211096394.6

(22)申请日 2022.09.06

(71)申请人 中国人民解放军国防科技大学
地址 210007 江苏省南京市秦淮区后标营18号

(72)发明人 张骁雄 丁松 丁鲲 李明浩 田昊 杨琴琴 张明星

(74)专利代理机构 江苏瑞途律师事务所 32346
专利代理师 韦超峰

(51)Int.Cl.
G06Q 50/26(2012.01)
G06Q 10/0637(2023.01)
G06F 30/27(2020.01)
G06N 3/092(2023.01)
G06F 111/04(2020.01)

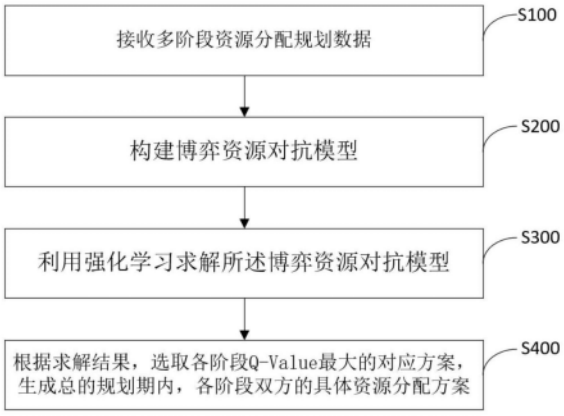
权利要求书3页 说明书9页 附图2页

(54)发明名称

基于强化学习的多阶段资源攻防分配方法及系统

(57)摘要

本发明公开了一种基于强化学习的多阶段资源攻防分配方法及系统，属于装备发展规划技术领域。所述方法包括接收多阶段资源分配规划数据；构建博弈资源对抗模型；利用强化学习求解所述博弈资源对抗模型；根据求解结果，选取各阶段Q-Value最大的对应方案，生成总的规划期内，各阶段双方的具体资源分配方案。所述系统基于上述方法进行多阶段资源攻防分配。本发明考虑了多阶段的资源分配问题，生成整个周期内的最优资源分配策略，能够为多阶段的攻防博弈资源分配提供辅助决策，使资源效益最大化。



CN 115619607 A

1.一种基于强化学习的多阶段资源攻防分配方法,其特征在于,包括:

接收多阶段资源分配规划数据;

构建博弈资源对抗模型;

利用强化学习求解所述博弈资源对抗模型;

根据求解结果,选取各阶段Q-Value最大的对应方案,生成总的规划期内,各阶段双方的具体资源分配方案。

2.根据权利要求1所述的方法,其特征在于,所述的博弈资源对抗模型包括:

在任一阶段t,存在如下资源约束:

$$n_t \times c \le \sum_{m=1}^{M} y_{t1}^{m}$$

其中,n_t表示该阶段进攻者共发动的侦察次数;c表示进攻者发动一次侦查行为所需消耗资源;M表示进攻者个数;m表示进攻者的指标;y_{t1}^{m}表示第m个进攻者在t阶段用于识别伪装目标的资源;

在n_t次侦查后剩余伪装目标个数为:

$$J_t = (H_t + J_{t-1}) \times (1-w)^{n_t}$$

其中,H_t表示防守者在t阶段新部署的伪装目标数目;J_t表示防守者在t阶段未被识别的伪装目标数目;w表示各阶段伪装目标被识别出来的概率;

在识别出K_t个伪装目标的情形下,防守者的真实目标在阶段t被击毁的概率:

$$p_t = \frac{(\sum_{m=1}^{M} y_{t2}^{m} / (J_t+1))^{\beta}}{(\sum_{m=1}^{M} y_{t2}^{m} / (J_t+1))^{\beta} + (x_{t2})^{\beta}}$$

其中,y_{t2}^{m}表示第m个进攻者在t阶段用于攻击剩余目标的资源;x_{t2}表示防守者在t阶段用于加强真实目标的资源;β表示攻防双方在实施打击中的对抗强度指标;

若真实目标未被击毁,则其在当前阶段t发挥效益为:

$v_t = \sigma * t$

其中,σ表示真实目标发挥价值增长率。

3.根据权利要求2所述的方法,其特征在于,所述的利用强化学习求解所述博弈资源对抗模型的方法包括:

在每个阶段,基于当前攻防双方的资源剩余情况,更新当前阶段的解空间;

通过序列博弈理论求得各阶段下的纳什均衡解;

基于上个阶段优化过程获得的解,利用最优Q-Value选择一个资源分配方案,并更新当前阶段下选择该资源分配方案解的Q-Value。

4.根据权利要求4所述的方法,其特征在于,在获取最优Q-Value时,需要构建回报函数,所述回报函数为:

$$R_t = (1-p_t) \times v_t + \sum_{x'} (1-p_{t+1}) \times v_{t+1} / N_{x'}$$

其中,p_t为t阶段成功击毁真实目标的概率;$1-p_t$为在t时刻真实目标存活的概率,v_t为真实目标t时刻发挥的价值,x'为防守者在下一阶段的可能方案,$N_{x'}$为下一阶段所有可能的防守方部署方案个数,p_{t+1}为防守者采取方案x'后的击毁概率。

5.根据权利要求4所述的方法,其特征在于,更新当前阶段下选择该资源分配方案解的Q-Value的方法为

$$Q(S_t,a_t)\leftarrow(1-\alpha)\times Q(S_t,a_t)+\alpha\left[R_t+\gamma\max_{PS}Q(S_{t+1},a_{t+1})\right]$$

其中,$Q(S_t,a_t)$代表在状态S_t下选择方案a_t的Q-Value,α是学习率,γ代表折扣因素。

6.一种基于强化学习的多阶段资源攻防分配系统,其特征在于,包括:

数据接收模块,其被配置为接收多阶段资源分配规划数据;

模型构建模块,其被配置为用于构建博弈资源对抗模型;

求解模块,其被配置为利用强化学习求解所述博弈资源对抗模型;

方案生成模块,其被配置为根据求解结果,选取各阶段Q-Value最大的对应方案,生成总的规划期内,各阶段双方的具体资源分配方案。

7.根据权利要求6所述的系统,其特征在于,所述的博弈资源对抗模型包括:

在任一阶段t,存在如下资源约束:

$$n_t\times c\le\sum_{m=1}^{M}y_{t1}^{m}$$

其中,n_t表示该阶段进攻者共发动的侦察次数;c表示进攻者发动一次侦查行为所需消耗资源;M表示进攻者个数;m表示进攻者的指标;y_{t1}^{m}表示第m个进攻者在t阶段用于识别伪装目标的资源;

在n_t次侦查后剩余伪装目标个数为:

$$J_t=(H_t+J_{t-1})\times(1-w)^{n_t}$$

其中,H_t表示防守者在t阶段新部署的伪装目标数目;J_t表示防守者在t阶段未被识别的伪装目标数目;w表示各阶段伪装目标被识别出来的概率;

在识别出K_t个伪装目标的情形下,防守者的真实目标在阶段t被击毁的概率:

$$p_t=\frac{(\sum_{m=1}^{M}y_{t2}^{m}/(J_t+1))^{\beta}}{(\sum_{m=1}^{M}y_{t2}^{m}/(J_t+1))^{\beta}+(x_{t2})^{\beta}}$$

其中,y_{t2}^{m}表示第m个进攻者在t阶段用于攻击剩余目标的资源;x_{t2}表示防守者在t阶段用于加强真实目标的资源;β表示攻防双方在实施打击中的对抗强度指标;

若真实目标未被击毁,则其在当前阶段t发挥效益为:

$v_t=\sigma*t$

其中,σ表示真实目标发挥价值增长率。

8.根据权利要求7所述的系统,其特征在于,所述的利用强化学习求解所述博弈资源对抗模型的方法包括:

在每个阶段,基于当前攻防双方的资源剩余情况,更新当前阶段的解空间;

通过序列博弈理论求得各阶段下的纳什均衡解;

基于上个阶段优化过程获得的解,利用最优Q-Value选择一个资源分配方案,并更新当前阶段下选择该资源分配方案解的Q-Value。

9.根据权利要求8所述的系统,其特征在于,在获取最优Q-Value时,需要构建回报函数,所述回报函数为:

$$R_t = (1-p_t)\times v_t + \sum_{x'}(1-p_{t+1})\times v_{t+1} / N_{x'}$$

其中，p_t为t阶段成功击毁真实目标的概率；$1-p_t$为在t时刻真实目标存活的概率，v_t为真实目标t时刻发挥的价值，x'为防守者在下一阶段的可能方案，$N_{x'}$为下一阶段所有可能的防守方部署方案个数，p_{t+1}为防守者采取方案x'后的击毁概率。

10.根据权利要求9所述的系统，其特征在于，更新当前阶段下选择该资源分配方案解的Q-Value的方法为

$$Q(S_t,a_t) \leftarrow (1-\alpha)\times Q(S_t,a_t) + \alpha\left[R_t + \gamma \max_{PS} Q(S_{t+1},a_{t+1})\right]$$

其中，$Q(S_t,a_t)$代表在状态S_t下选择方案a_t的Q-Value，α是学习率，γ代表折扣因素。

基于强化学习的多阶段资源攻防分配方法及系统

技术领域

[0001] 本发明属于装备发展规划技术领域，具体涉及一种基于强化学习的多阶段资源分配方法及系统。

背景技术

[0002] 目前，国土安全防御方面的研究已引起广泛关注，主要聚焦如何利用有限资源实现效益最大化。通常来说，防守方需要合理部署防护资源，保护自身重要目标；而进攻方合理部署资源实施打击，旨在摧毁重要目标已提出不同博弈模型，以丰富现实中各种可能情形下的资源分配问题。针对上述攻防问题，不同研究人员基于博弈论展开了广泛研究，以更好表征和度量攻防双方的对抗行为。

[0003] 很多研究中已提出不同博弈模型，以丰富现实中各种情形下的资源分配问题。。然而目前大部分研究局限于考虑单阶段单进攻者的情况，而多阶段的资源分配在现实中同样常见，但现有研究中鲜有研究考虑多阶段多进攻者情况下的资源分配情况。通常在多阶段情况发生时，仅是将单阶段的资源分配进行简单叠加。但随着规划周期和阶段的增多，现有的方法已经难以满足需求。

发明内容

[0004] 技术问题：本发明提供一种考虑多阶段情形的基于强化学习的多阶段资源攻防分配方法及系统，从而为多阶段的攻防博弈资源分配提供辅助决策，使资源效益最大化。

[0005] 技术方案：第一方面，本发明提供一种基于强化学习的多阶段资源攻防分配方法，其特征在于，包括：

[0006] 接收多阶段资源分配规划数据；

[0007] 构建博弈资源对抗模型；

[0008] 利用强化学习求解所述博弈资源对抗模型；

[0009] 根据求解结果，选取各阶段Q-Value最大的对应方案，生成总的规划期内，各阶段双方的具体资源分配方案。

[0010] 进一步地，所述的博弈资源对抗模型包括：

[0011] 在任一阶段t，存在如下资源约束：

[0012] $$n_t \times c \le \sum_{m=1}^{M} y_{t1}^{m}$$

[0013] 其中，n_t表示该阶段进攻者共发动的侦察次数；c表示进攻者发动一次侦查行为所需消耗资源；M表示进攻者个数；m表示进攻者的指标；y_{t1}^{m}表示第m个进攻者在t阶段用于识别伪装目标的资源；

[0014] 在n_t次侦查后剩余伪装目标个数为：

[0015] $$J_t = (H_t + J_{t-1}) \times (1-w)^{n_t}$$

[0016] 其中，H_t表示防守者在t阶段新部署的伪装目标数目；J_t表示防守者在t阶段未被识

别的伪装目标数目;w表示各阶段伪装目标被识别出来的概率;

[0017] 在识别出K_t个伪装目标的情形下,防守者的真实目标在阶段t被击毁的概率:

[0018] $$p_t=\frac{(\sum_{m=1}^{M}y_{t2}^{m}/(J_t+1))^{\beta}}{(\sum_{m=1}^{M}y_{t2}^{m}/(J_t+1))^{\beta}+(x_{t2})^{\beta}}$$

[0019] 其中,y_{t2}^{m}表示第m个进攻者在t阶段用于攻击剩余目标的资源;x_{t2}表示防守者在t阶段用于加强真实目标的资源;β表示攻防双方在实施打击中的对抗强度指标;

[0020] 若真实目标未被击毁,则其在当前阶段t发挥效益为:

[0021] $v_t=\sigma*t$

[0022] 其中,σ表示真实目标发挥价值增长率。

[0023] 进一步地,所述的利用强化学习求解所述博弈资源对抗模型的方法包括:

[0024] 在每个阶段,基于当前攻防双方的资源剩余情况,更新当前阶段的解空间;

[0025] 通过序列博弈理论求得各阶段下的纳什均衡解;

[0026] 基于上个阶段优化过程获得的解,利用最优Q-Value选择一个资源分配方案,并更新当前阶段下选择该资源分配方案解的Q-Value。

[0027] 进一步地,在获取最优Q-Value时,需要构建回报函数,所述回报函数为:

[0028] $$R_t=(1-p_t)\times v_t+\sum_{x'}(1-p_{t+1})\times v_{t+1}/N_{x'}$$

[0029] 其中,p_t为t阶段成功击毁真实目标的概率;$1-p_t$为在t时刻真实目标存活的概率,v_t为真实目标t时刻发挥的价值,x'为防守者在下一阶段的可能方案,$N_{x'}$为下一阶段所有可能的防守方部署方案个数,p_{t+1}为防守者采取方案x'后的击毁概率。

[0030] 进一步地,更新当前阶段下选择该资源分配方案解的Q-Value的方法为

[0031] $$Q(S_t,a_t)\leftarrow(1-\alpha)\times Q(S_t,a_t)+\alpha\left[R_t+\gamma\max_{PS}Q(S_{t+1},a_{t+1})\right]$$

[0032] 其中,$Q(S_t,a_t)$代表在状态S_t下选择方案a_t的Q-Value,α是学习率,γ代表折扣因素。

[0033] 第二方面,本发明提供一种基于强化学习的多阶段资源攻防分配系统,包括:

[0034] 数据接收模块,其被配置为接收多阶段资源分配规划数据;

[0035] 模型构建模块,其被配置为用于构建博弈资源对抗模型;

[0036] 求解模块,其被配置为利用强化学习求解所述博弈资源对抗模型;

[0037] 方案生成模块,其被配置为根据求解结果,选取各阶段Q-Value最大的对应方案,生成总的规划期内,各阶段双方的具体资源分配方案。

[0038] 进一步地,所述的博弈资源对抗模型包括:

[0039] 在任一阶段t,存在如下资源约束:

[0040] $$n_t\times c\le\sum_{m=1}^{M}y_{t1}^{m}$$

[0041] 其中,n_t表示该阶段进攻者共发动的侦察次数;c表示进攻者发动一次侦查行为所需消耗资源;M表示进攻者个数;m表示进攻者的指标;y_{t1}^{m}表示第m个进攻者在t阶段用于识别伪装目标的资源;

[0042] 在n_t次侦查后剩余伪装目标个数为:

[0043] $J_t = (H_t + J_{t-1}) \times (1-w)^{n_t}$

[0044] 其中，H_t表示防守者在t阶段新部署的伪装目标数目；J_t表示防守者在t阶段未被识别的伪装目标数目；w表示各阶段伪装目标被识别出来的概率；

[0045] 在识别出K_t个伪装目标的情形下，防守者的真实目标在阶段t被击毁的概率：

[0046] $$p_t = \frac{(\sum_{m=1}^{M} y_{t2}^m / (J_t+1))^{\beta}}{(\sum_{m=1}^{M} y_{t2}^m / (J_t+1))^{\beta} + (x_{t2})^{\beta}}$$

[0047] 其中，y_{t2}^m表示第m个进攻者在t阶段用于攻击剩余目标的资源；x_{t2}表示防守者在t阶段用于加强真实目标的资源；β表示攻防双方在实施打击中的对抗强度指标；

[0048] 若真实目标未被击毁，则其在当前阶段t发挥效益为：

[0049] $v_t = \sigma * t$

[0050] 其中，σ表示真实目标发挥价值增长率。

[0051] 进一步地，所述的利用强化学习求解所述博弈资源对抗模型的方法包括：

[0052] 在每个阶段，基于当前攻防双方的资源剩余情况，更新当前阶段的解空间；

[0053] 通过序列博弈理论求得各阶段下的纳什均衡解；

[0054] 基于上个阶段优化过程获得的解，利用最优Q-Value选择一个资源分配方案，并更新当前阶段下选择该资源分配方案解的Q-Value。

[0055] 进一步地，在获取最优Q-Value时，需要构建回报函数，所述回报函数为：

[0056] $$R_t = (1-p_t) \times v_t + \sum_{x'} (1-p_{t+1}) \times v_{t+1} / N_{x'}$$

[0057] 其中，p_t为t阶段成功击毁真实目标的概率；$1-p_t$为在t时刻真实目标存活的概率，v_t为真实目标t时刻发挥的价值，x'为防守者在下一阶段的可能方案，$N_{x'}$为下一阶段所有可能的防守方部署方案个数，p_{t+1}为防守者采取方案x'后的击毁概率。

[0058] 进一步地，更新当前阶段下选择该资源分配方案解的Q-Value的方法为

[0059] $$Q(S_t, a_t) \leftarrow (1-\alpha) \times Q(S_t, a_t) + \alpha \left[R_t + \gamma \max_{PS} Q(S_{t+1}, a_{t+1}) \right]$$

[0060] 其中，$Q(S_t, a_t)$代表在状态S_t下选择方案a_t的Q-Value，α是学习率，γ代表折扣因素。

[0061] 本发明与现有技术相比，研究了多阶段下考虑多进攻者威胁的资源分配，并将将强化学习应用于多阶段的资源分配，借鉴博弈论均衡解的概念，模型在各阶段的分配中，防守者旨在最大化目标效益，而进攻者旨在最大化摧毁真实目标。采用强化学习可有效对多阶段问题进行水平搜索，形成任意阶段的策略规则，从而有效保证决策结果在整个阶段的最优性。本发明能够生成整个周期内的最优资源分配策略，能够为多阶段的攻防博弈资源分配提供辅助决策，使资源效益最大化。

附图说明

[0062] 图1为本发明实施例中基于强化学习的多阶段资源攻防分配方法的流程图；

[0063] 图2为本发明的示例中各阶段双方资源分配结果示意图；

[0064] 图3为本发明的示例中是否采取强化学习下防守者各阶段期望效益图；

[0065] 图4为本发明的示例中进攻者是否合作情形下防守者各阶段期望效益图。

具体实施方式

[0066] 下面结合实施例和说明书附图对本发明作进一步的说明。

[0067] 首先,对实施例中涉及的符号进行解释说明:

[0068] M:进攻者个数;

[0069] m＝{1,…,M}:进攻者的指标;

[0070] T:阶段个数;

[0071] t＝{1,…,T}:阶段的指标;

[0072] X:防守者资源约束;

[0073] Y_m:第m个进攻者的资源约束,m＝{1,…,M};

[0074] x_{t1}:防守者在t阶段用于部署伪装目标的资源;

[0075] x_{t2}:防守者在t阶段用于加强真实目标的资源;

[0076] s:部署一个伪装目标所需消耗资源;

[0077] c:进攻者发动一次侦查行为所需消耗资源;

[0078] W:各阶段伪装目标被识别出来的概率;

[0079] H_t:防守者在t阶段新部署的伪装目标数目;

[0080] K_t:防守者在t阶段被识别的伪装目标数目;

[0081] J_t:防守者在t阶段未被识别的伪装目标数目;

[0082] p_t:t阶段成功击毁真实目标的概率;

[0083] y_{t1}^m:第m个进攻者在t阶段用于识别伪装目标的资源;

[0084] y_{t2}^m:第m个进攻者在t阶段用于攻击剩余目标的资源;

[0085] δ:真实目标发挥价值增长率;

[0086] v_t:防守者真实目标在t阶段发挥的效用;

[0087] β:攻防双方在实施打击中的对抗强度指标;

[0088] g:防守者在各阶段的资源分配比例;

[0089] h:进攻者在各阶段的资源分配比例。

[0090] 图1示出了本发明所提出的基于强化学习的多阶段资源攻防分配方法的流程图。结合图1所示,本发明包括如下步骤:

[0091] 步骤S100:接收多阶段资源分配规划数据。可能需要的数据包括:进攻者个数、阶段个数、防守者资源约束、进攻者资源约束、部署一个伪装目标所需资源、识别一个伪装目标所需资源、伪装目标被识别的概率、真实目标价值增长率、攻防双方对抗强度指标、防守者各阶段资源分配比例和进攻者各阶段资源分配比例等。

[0092] 步骤S200:构建博弈资源对抗模型。

[0093] 本发明重点聚焦一个防守者与多个进攻者之间的序列博弈中,多阶段情形下的防护资源分配问题:有限的资源条件下,防守者如何将防护措施按阶段合理分配以最大程度发挥真实目标的效益?进攻者集体在观察到防守者的动作后,如何相互配合选择各自的进攻策略,以获取集体最大的收益?相比于传统的序列对抗,多阶段资源分配,更加突出时间维度,并非将单阶段下的资源分配进行简单叠加。其特点在于之前阶段的策略选择直接影

响后续阶段的解空间,阶段之间相互紧密关联。本发明基于如下基本假设:

[0094] (1)各阶段中,防守者可分配资源用于加固真实目标,也可部署伪装目标以迷惑进攻者;

[0095] (2)各阶段中,各进攻者可分配资源识别伪装目标,也可在剩余目标之间平均分配资源实施无差别打击;

[0096] (3)进攻者实施打击成功概率取决于双方在该目标上投入的资源,下文中采用"对抗函数"进行度量;

[0097] (4)进攻者各阶段可分配一定资源(比如发动一次侦查行为)以一定概率识别出伪装目标,也可发动多次侦查,每次侦查行为相对独立;

[0098] (5)一次成功的攻击可摧毁真实目标,而失败的攻击不会对真实目标产生损伤;

[0099] (6)攻、防双方在各阶段存在资源约束,且当前阶段的资源不累加到下个阶段使用;

[0100] (7)进攻者之间协同合作,进攻者的资源分配可以叠加使用,最大化集体的效益价值;

[0101] (8)进攻者在各阶段选择最大化自身利益的资源分配策略,即最小化真实目标发挥的效用;

[0102] (9)防守者旨在最大化真实目标在多个阶段中发挥的累积效用,进攻者旨在最小化该目标在各阶段中发挥的累积效用。

[0103] 在序列博弈中,各阶段攻防双方存在两次交互行为。第一次交互行为聚焦于防守者部署伪装目标迷惑进攻者,而进攻者分配一定资源识别伪装目标。第二次交互行为聚焦于防守者部署资源加强真实目标,而进攻者分配一定资源攻击剩余未被识别目标。

[0104] 在任一阶段t,由于进攻者无法准确识别真实目标和伪装目标,假设在该阶段进攻者共发动了n_t次侦查行为($n_t \geq 0$),则当前阶段存在如下资源约束:

[0105] $$n_t \times c \leq \sum_{m=1}^{M} y_{t1}^{m} \qquad (1)$$

[0106] 其中,$\sum_{m=1}^{M} y_{t1}^{m}$表示t阶段进攻者用于识别伪装目标资源总和。

[0107] 假设任一阶段新部署H_t个伪装目标,上一阶段剩余J_{t-1}个未被成功识别伪装目标。由于各阶段的多次侦查行为相对独立,则在n_t次侦查后剩余伪装目标个数为:

[0108] $$J_t = (H_t + J_{t-1}) \times (1-w)^{n_t} \qquad (2)$$

[0109] 其中,H_t+J_{t-1}为发动侦查行为之前当前阶段所有的伪装目标个数之和,

[0110] $J_t \geq 0$,$J_0 = 0$。

[0111] 在任一阶段t,进攻者识别完伪装目标后,在剩余未被识别的伪装目标Jt和真实目标之间实施无差别攻击。防守者在该阶段分配资源x_{t2}加固真实目标,而攻击者在剩余目标之间平均分配资源,则分配在真实目标上的攻击资源为$\sum_{m=1}^{M} y_{t2}^{m} / (J_t+1)$。因此,在识别出$K_t$个伪装目标的情形下,防守者的真实目标在阶段t被击毁的概率采用如下竞赛函数来表示:

[0112] $$p_t = \frac{(\sum_{m=1}^{M} y_{t2}^{m} / (J_t+1))^{\beta}}{(\sum_{m=1}^{M} y_{t2}^{m} / (J_t+1))^{\beta} + (x_{t2})^{\beta}} \qquad (3)$$

[0113] 此处,β为博弈竞赛强度。当β=0时,双方对博弈结果影响相同。0<β<1时候,投入较

少资源的选手具备更大的优势；该β＝1时，双方对抗结果与双方投入的资源成正比；当β趋于无穷大时候，会出现“获胜者赢取一切”的现象。

[0114] 若真实目标未被击毁，则其在当前阶段t发挥效益为

[0115] $v_t=\sigma*t$ (4)

[0116] 防守者旨在最大化各阶段的累加期望效用，而不同进攻者旨在最小化真实目标的累加期望效用。同时，进攻者可以相互合作，共享各自的资源来对防守者实施攻击行为，若不合作，则各自相互独立，每名进攻者仅根据自身的资源约束进行决策。

[0117] 步骤S300：利用强化学习求解所述博弈资源对抗模型。

[0118] 基于强化学习的多阶段资源分配问题求解步骤如下：

[0119] (1)在每个阶段，基于当前攻防双方的资源剩余情况，更新当前阶段的解空间；

[0120] (2)序列博弈理论求得各阶段下的纳什均衡解。即防守者先分配资源用于加固真实目标和部署伪装目标；进攻者分配资源识别伪装目标和对剩余目标实施无差别打击；

[0121] (3)基于上个阶段优化过程获得的解，采用随机探索或者利用最优Q-Value选择一个资源分配方案，并更新当前阶段下选择该资源分配方案解的Q-Value；

[0122] (4)迭代上述步骤，直到达到停止标准。

[0123] 在每一次的迭代计算中，通过计算当前阶段序列博弈问题生成资源分配方案。各个阶段存在资源约束情形下，防守者对应不同部署伪装目标个数以及加强真实目标的策略选择，每一个防守方案下进攻者也存在一个最优进攻选择，从所有选择组合中选择Value值最高的资源分配方案，即Q-Learning行为。在t＝0处进行选择之后，然后根据所选动作生成下一个状态S_{t+1}。

[0124] 在步骤(3)中涉及回报函数的构建，是衡量和计算资源分配方案Q-Value的重要依据。对于防守者，假设当前阶段S_t选择某个资源分配方案a_t的回报值R_t可以通过下式计算：

[0125] $$R_t=(1-p_t)\times v_t+\sum_{x'}(1-p_{t+1})\times v_{t+1}/N_{x'} \quad (7)$$

[0126] 其中，$1-p_t$为在t时刻真实目标存活的概率，v_t为真实目标t时刻发挥的价值。本发明以真实目标发挥的期望效益来衡量当前资源分配方案的回报值。x’为防守者在下一阶段的可能方案，$N_{x'}$为下一阶段所有可能的防守方部署方案个数，p_{t+1}为防守者采取方案x’后的击毁概率。传统的博弈论求解方法关注当前阶段的期望收益，即式子中$(1-p_t)\times v_t$部分，来衡量资源分配效果，本发明回报函数的设定同时不仅仅考虑当前阶段的收益，同时考虑下一阶段的期望收益。通过两者的加和来衡量方案的优劣。

[0127] 一旦计算出选中的资源分配方案对应的回报值R_t，则可通过如下公式对Q-Value进行更新：

[0128] $$Q(S_t,a_t)\leftarrow(1-\alpha)\times Q(S_t,a_t)+\alpha\left[R_t+\gamma\max_{PS}Q(S_{t+1},a_{t+1})\right] \quad (8)$$

[0129] 其中，$Q(S_t,a_t)$代表在状态S_t下，选择方案a_t的Q-Value，α∈[0,1]是学习率，用于决定新信息被采用的程度，α＝0代表不学习新信息，α＝1代表只学习最近更新的信息。通常随机环境下选择一个较小。

[0130] 步骤S400：根据求解结果，选取各阶段Q-Value最大的对应方案，生成总的规划期内，各阶段双方的具体资源分配方案。求解后可生成不同阶段的Q-Value矩阵，对应不同的

攻防资源分配方案，选取各阶段Q-Value最大的对应方案，生成总的规划期内，各阶段双方的具体资源分配方案。

[0131] 进一步地，为了验证本发明所提供方法的有效性，提供一个仿真示例进行验证。该示例中，参数设置如下：

[0132] (1)进攻者个数：假定存在3个进攻者，M＝3；

[0133] (2)阶段个数：假定存在3个攻防阶段，T＝3；

[0134] (3)防守者资源约束：X＝25亿元；

[0135] (4)进攻者资源约束：Y＝[8 12 10]亿元；

[0136] (5)部署一个伪装目标所需资源：s＝1.6亿元；

[0137] (6)识别一个伪装目标所需资源：c＝1.1亿元；

[0138] (7)伪装目标被识别的概率：w＝0.3；

[0139] (8)真实目标价值增长率：$\delta=5$；

[0140] (9)攻防双方对抗强度指标：β＝0.5；

[0141] (10)防守者各阶段资源分配比例：g＝[0.3 0.5 0.2]；

[0142] (11)进攻者各阶段资源分配比例：h＝[0.2 0.6 0.20.3 0.5 0.20.4 0.3 0.3]。

[0143] 基于上述方法，运行后可生成不同阶段的Q值矩阵，对应不同的攻防资源分配方案。选取各阶段Q值最大的对应方案，生成总的规划期内，各阶段双方的具体资源分配方案如图2所示。

[0144] 上图分别展示了攻防双方在三个阶段上对资源的分配情况。其中，蓝色柱体表示防守者/进攻者在部署/识别伪装目标上的资源分配情况，而黄色柱体表示防守者/进攻者在巩固真实目标/攻击真实目标上的资源分配情况。

[0145] 由图2知，防守者在第一阶段消耗3.2个单位资源部署了2个伪装目标，而三名进攻者联合分别分配0.545、0.737和0.918个单位资源，发动2次侦查行为用于识别伪装目标，攻防双方都将该阶段剩余资源全部用于防护/攻击真实目标。第二阶段，防守者仍然消耗1.6个单位资源部署1个伪装目标迷惑进攻者，而三名进攻者分别分配1.594、2.218和0.588个单位资源，共发起4轮侦查行为用于识别伪装目标，之后将当前阶段全部资源都用于在剩余目标之间实行无差别攻击。在第三阶段，防守者仍然消耗1.6个单位资源部署1个伪装目标，而三名进攻者分别分配0.336、0.277和0.487个单位资源，实施1次侦查行为识别伪装目标，之后将全部资源用于攻击所有剩余目标。

[0146] 通过此资源分配方案，防守者真实目标在三个阶段的存活概率分别为0.52，0.57和0.52，且真实目标在整个阶段发挥的期望效益值累计达到7.91。

[0147] 为验证参数对模型结果的影响，保持其他相关参数设置不变，分别改变目标价值增长率指标以及攻防对抗强度指标，对不同参数下的方案结果进行对比分析。

[0148] 分别设置真实目标价值增长率为5、10、15、20和25，生成5种情形下的资源分配方案。

[0149] 表1不同价值增长率下攻防资源分配对比

[0150]

参数	各阶段部署伪装目标/发动侦查次数		期望效用
б=5	防守者	2，1，1	7.91
	进攻者	2，4，1	
б=10	防守者	2，1，1	15.82
	进攻者	2，4，1	
б=15	防守者	2，1，1	23.74
	进攻者	2，4，1	
б=20	防守者	2，1，1	31.65
	进攻者	2，4，1	
б=25	防守者	2，1，1	39.56
	进攻者	2，4，1	

[0151] 表1展示了不同真实目标价值增长率下攻防双方在各阶段的资源分配部署。其中，选取部署伪装目标/识别伪装目标的侦查行为次数来衡量各阶段双方的资源分配，剩余各阶段资源即全部用于加强真实目标和攻击真实目标。结果发现各情形下双方的最优资源分配结果一致，即防守者分别在三个阶段部署2、1和1个伪装目标，而进攻者分别在各个阶段实施2、4和1次侦查行为。防守者的期望效用随着真实目标价值的增大而增大。研究结果表明，攻防双方的资源分配策略不随真实目标的价值而改变。

[0152] 分别设置攻防对抗强度指标为0.1、0.5、1、5和10，生成5种情形下的资源分配方案。

[0153] 表2不同价值增长率下攻防资源分配对比

[0154]

参数	各阶段部署伪装目标/发动侦查次数		各阶段真实目标存活概率	期望效用
$\beta=0.1$	防守者	2，1，1	0.50，0.51，0.50	7.07
	进攻者	2，4，1		
$\beta=0.5$	防守者	2，1，1	0.52，0.57，0.52	7.91
	进攻者	2，4，1		
$\beta=1$	防守者	2，1，1	0.55，0.63，0.54	9.01
	进攻者	2，4，1		
$\beta=5$	防守者	2，3，1	0.73，0.90，0.75	17.55
	进攻者	2，5，1		
$\beta=10$	防守者	1，4，1	0.89，0.94，0.93	24.62
	进攻者	1，5，1		

[0155] 表2结果表明，对抗强度参数的变化导致攻防双方的资源分配结果略有差异。当β不超过1时候，攻防双方在三种情形下的资源分配相同，但效用随着β的增大而增大，主要得益于真实目标在各阶段的存活率不断增大。相比于其他阶段的资源分配，当β为5的时候，防守者分别在三个阶段部署2、3和1个伪装目标，而进攻者分别在三个阶段实施2、5和1次侦查行为。当β为10的时候，防守者分别在三个阶段部署1、4和1个伪装目标，而进攻者分别在三个阶段实施1、5和1次侦查行为。相比与之前行为，当β增大时候，防守者和进攻者分别在第二阶段加大了部署伪装目标和侦查行为的力度。当β不断增大时候，防守者不断调整自身在各阶段的资源分配情况，使得真实目标在各阶段的存活概率不断增大。

[0156] 进一步地，为验证是否采用Q学习对防守者资源分配策略的影响，继续开展对比实验。采用强化学习策略依据Q函数选取策略时，不仅仅考虑当前阶段最优，同时考虑下一阶段。不采取强化学习策略时，防守者只选取当前阶段的纳什均衡最优解。图3给出了是否采取强化学习策略下防守者真实目标在各阶段发挥的期望效益结果。

[0157] 由图3知，第一阶段考虑强化学习策略的分配方案所产生的期望收益略低于不考

虑强化学习策略对应的收益，原因可能在于真实目标发挥的效益随着时间不断增大，因此防守者期望真实目标在下一阶段能发挥更大效益。对比发现第二阶段和第三阶段下的期望效益，采用强化学习策略下真实目标发挥期望效益要高于不采取强化学习。且利用Q学习策略生成的资源方案，可以保证真实目标在整个阶段发挥的期望效益累计值更优。

[0158] 上文中各场景实验均假设不同的攻击者相互合作，即以最大化群体的效益为目标，共享支配各自的资源分配。下文拟假设各位进攻者不相互合作，即各阶段的进攻资源分配选择仅从最大化自身的效益出发。图4给出了进攻者是否合作策略下防守者的真实目标在各阶段发挥的期望效益结果。

[0159] 由图4知，第一阶段两种策略的期望收益相同，在后面的第二和第三两个阶段，进攻者不合作情形下防守者真实目标发挥的期望效益都相对更高。即进攻者不合作对防守者更加有利，真实目标具备更高的存活概率且可以发挥更高的效益价值。对比两种情形下的资源分配方案，防守者的资源策略保持不变，但任一进攻者都不再分配资源去识别伪装目标。此分配策略的部分原因是任一单独进攻者在各阶段的资源有限，可能无法有效支撑用于识别伪装目标，因此单个进攻者更倾向于将有限的资源用于实施攻击。

[0160] 进一步地，本发明提供一种基于强化学习的多阶段资源攻防分配系统，该系统主要是基于上文所提出的任一基于强化学习的多阶段资源攻防分配方法进行多阶段资源攻防分配。该系统包括：数据接收模块、模型构建模块、求解模块和方案生成模块。其中，数据接收模块被配置为接收多阶段资源分配规划数据；模型构建模块被配置为用于构建博弈资源对抗模型；求解模块被配置为利用强化学习求解所述博弈资源对抗模型；方案生成模块被配置为根据求解结果，选取各阶段Q-Value最大的对应方案，生成总的规划期内，各阶段双方的具体资源分配方案。对于各模块具体如何实现其功能，与所提供的方法的相应步骤相对应，此处不再赘述。

[0161] 资源分配是攻防博弈中十分重要的现实问题，针对多阶段的攻防资源分配，本发明重点研究了多阶段下考虑多进攻者威胁的资源分配。本发明将强化学习应用于多阶段的资源分配中，借鉴博弈论均衡解的概念，模型在各阶段的分配中，防守者旨在最大化目标效益，而进攻者旨在最大化摧毁真实目标。采用强化学习可有效对多阶段问题进行水平搜索，形成任意阶段的策略规则，从而有效保证决策结果在整个阶段的最优性。

[0162] 上述实施例仅是本发明的优选实施方式，应当指出：对于本技术领域的普通技术人员来说，在不脱离本发明原理的前提下，还可以做出若干改进和等同替换，这些对本发明权利要求进行改进和等同替换后的技术方案，均落入本发明的保护范围。

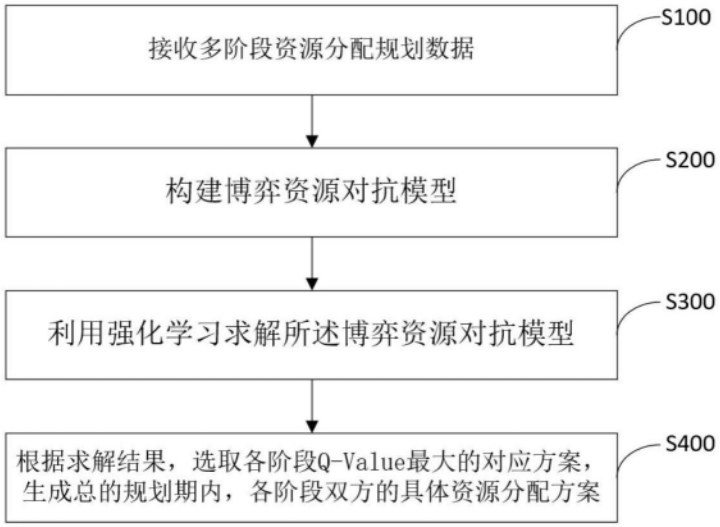

接收多阶段资源分配规划数据
S100
构建博弈资源对抗模型
S200
利用强化学习求解所述博弈资源对抗模型
S300
根据求解结果，选取各阶段Q-Value最大的对应方案，
生成总的规划期内，各阶段双方的具体资源分配方案
S400

图1

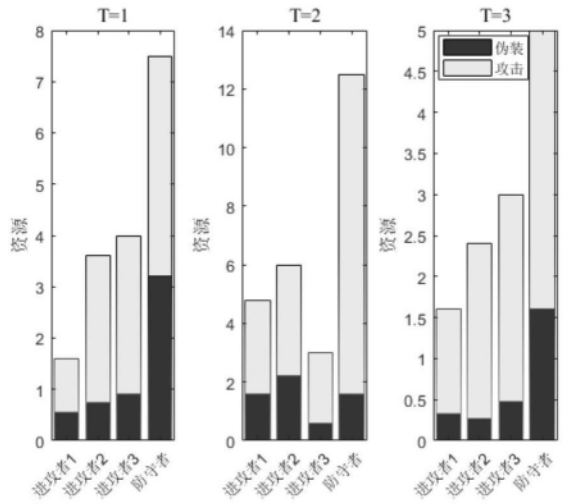

T=1
T=2
T=3
资源
伪装
攻击
进攻者1
进攻者2
进攻者3
防守者

图2

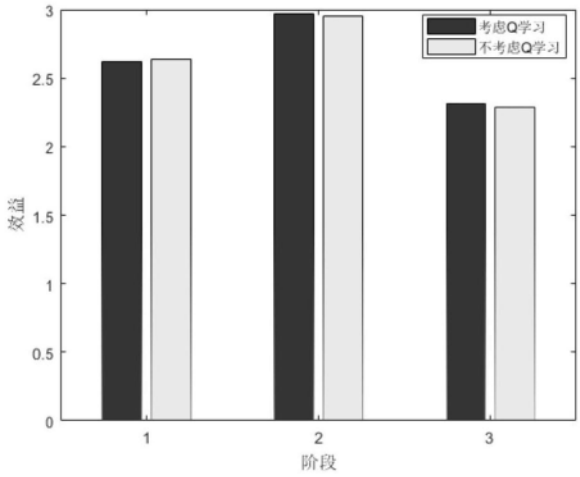

考虑Q学习
不考虑Q学习
效益
阶段
0
0.5
1
1.5
2
2.5
3
1
2
3

图3

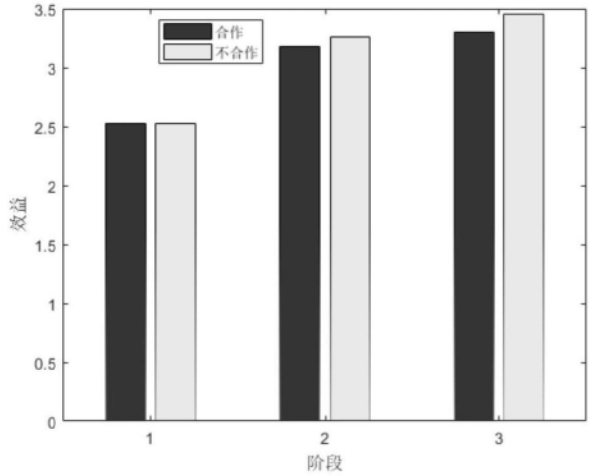

合作
不合作
效益
阶段
0
0.5
1
1.5
2
2.5
3
3.5
1
2
3

图4