文章编号: 1001-0920(2019)09-1917-06 DOI: 10.13195/j.kzyjc.2018.0082

基于改进Q-学习算法的多阶段群体决策模型

张 峰, 刘凌云†, 郭欣欣

(1. 河北大学 数学与信息科学学院，河北 保定 071002；
2. 河北省机器学习与计算智能重点实验室，河北 保定 071002)

摘 要: 多阶段群体决策问题是一类典型的动态群体决策问题,主要针对离散的确定状态下的最优群体决策问题求解. 但由于现实环境面临的大部分是不确定状态空间,甚至是未知环境空间(例如状态转移概率矩阵完全未知),为了寻求具有较高共识度的多阶段群体最优策略,决策者需要通过对环境的动态交互来获得进一步的信息. 针对该问题,利用强化学习技术,提出一种求解多阶段群体决策的最优决策算法,以解决在不确定状态空间下的多阶段群体决策问题. 结合强化学习中的Q-学习算法,建立多阶段群体决策Q-学习基本算法模型,并改进该算法的迭代过程,从中学习得到群体最优策略. 同时证明基于Q-学习得到的多阶段群体最优策略也是群体共识度最高的策略. 最后,通过一个计算实例说明算法的合理性及可行性.

关键词: 群体决策；多阶段群体决策；强化学习；Q-学习；群体共识；不确定性

中图分类号: TP181；TP39 **文献标志码:** A

A multi-stage group decision model based on improved Q-learning

ZHANG Feng, LIU Ling-yun†, GUO Xin-xin

(1. College of Mathematics and Information Science，Hebei University，Baoding 071002，China；2. Hebei Key Laboratory of Machine Learning and Computational Intelligence，Baoding 071002，China)

Abstract: The multi-stage group decision making problem is a typical sequential group decision making problem. It is normally utilized to find the optimal solution to the group decision problems in discrete deterministic environment. However, the real life environments faced by decision-makers are usually full of uncertainty, even unknown environments (with unknown state transition matrix). Therefore, it is essential for the decision-makers to obtain more information by interacting with the environment dynamically to achieve an optimal decision strategy with high consensus degree. Due to the advantage of reinforcement learning in handling the sequential decision-making problems, the classical reinforcement learning algorithm (Q-learning) is improved to discover the optimal solution of multi-stage group decision making problems under uncertain environment. Additionally, a theorem is proposed to show that the optimal group decision obtained by using the improved Q-learning algorithm is the group decision with the highest degree of group consensus. Finally, an illustrative example is presented to verify the rationality and feasibility of the proposed algorithm.

Keywords: group decision making；multi-stage group decision；reinforcement learning；Q-learning；group consensus；uncertainty

0 引 言

多阶段群体决策是指决策群体所要解决的问题包含多个决策阶段的一类群体决策,属于一类典型的动态群体决策问题[1]. 在多阶段群体决策问题中,决策群体需要在每个阶段作出相应的决策,各个阶段的群体决策形成一个连续的序列决策,这个连续的序列决策称为多阶段群体决策的策略. 在现实生活中,多阶段群体决策问题应用性非常广阔. 比如,项目投资一般情况下是由一系列不同阶段组成的,在企业进行风险投资项目评估时,通常是由多个专家组成的评审团对不同阶段的备选方案进行评价和选择,这样有利于项目管理者在每个阶段根据所得到的信息来判断下一阶段如何决策,多阶段群体决策的目标是使得群体在较高共识度的基础上获得整个过程的利益最大而不是阶段最优.

为了求解多阶段群体决策问题,彭怡等[2]建立了一种基于动态规划的多阶段群体决策模型,通过定义Pareto最优策略及绝对最优策略等概念,设计

收稿日期: 2018-01-16；**修回日期:** 2018-05-30.

基金项目: 国家自然科学基金项目(61672205)；河北省自然科学面上基金项目(F2017201020, F2018201115)；河北省教育厅青年基金项目(QN2015026, QN2017019).

责任编委: 刘宝碇.

†通讯作者. E-mail: 1058738029@qq.com.

并实现了求解Pareto最优策略的群体动态规划算法;郝晶晶等[3]提出了一类双重信息的多阶段群体决策方法来解决专家群体存在阶段差异的多阶段决策问题;Zhang[4]研究了在离散确定状态下的多阶段群体决策问题,将多阶段群体决策的问题与图的顶点集和边集相对应,根据图论原理建立数学模型,将在离散确定状态下的多阶段群体决策问题被转化为在图中有8个向量的多图的最长路径,并在此基础上给出了一种具有加权向量的最长路径问题的算法;Lu等[5]基于决策的横向和纵向信息,首先将决策者的判断集合在一个阶段,以获得阶段决策者的偏好和权重,并计算出群体的偏好,其次根据偏好距离模型计算出不同阶段决策者的满意偏好和阶段权重,最后利用偏好距离算法计算出优化决策矩阵和决策者的权重,得到最优组的偏好.

然而,以上针对多阶段的群体决策研究都是基于确定性的环境,而对于不确定环境下多阶段群体决策的研究,目前主要有一系列灰色决策方法和模糊语言决策方法. 张娜等[6]提出了一种多阶段灰色局势群体决策评价信息集结方法;Luo等[7]针对属性值为灰信息的决策问题,提出了一种基于灰信息的多阶段多属性风险型群决策方法;周声海[8]针对多阶段交互不确定性决策的动态过程,首先提出了一个迭代算子,进行聚集之间的冲突消解,使得各个决策成员的偏好达到一致性水平,然后考虑各个决策阶段之间的关系,通过引入马尔科夫链构建马尔科夫状态转移矩阵,从而进行多阶段交互式决策;马跃如等[9]针对物流合作伙伴选择过程中所面临的信息不完全性,引入模糊语言变量来描述决策者的评估信息,提出了多时段条件下动态物流联盟伙伴选择的模糊语言群体决策模型.

在现实生活中,问题的环境往往是完全不确定性的,甚至是未知的,比如某制药公司要开发一种新药品,制药过程包括研究开发、临床试验、生产及销售4个阶段,每一阶段有不同的备选方案,由于环境未知,选择不同的方案对每一阶段下环境的影响也未知.

求解不确定状态空间下的多阶段群体决策问题的难点在于:1) 如何在不确定状态下充分挖掘到有用的信息,使得决策群体作出的决策达到群体决策满意度最大;2) 决策群体如何在多个不确定阶段下进行交互,既能使得群体决策满意值最大,又使得群体共识度最高;3) 如何构建不确定状态空间下的多阶段群体决策,最终能求解得到全局最优决策. 为了解决以上问题,本文利用强化学习技术,寻求在不确定状态空间下多阶段群体决策的最优决策. 由于强化学习可以通过试错和动态环境交互而获得更多的信息,从而使得选取的动作能够获得环境累积回报最大,而最大的累积回报存在一个序列决策与之相对应,所以本文将多阶段性群体决策问题与强化学习联系起来,将强化学习技术运用到多阶段性群体决策中,以实现在不确定状态空间下求得多阶段群体决策的全局最优决策结果.

此外,群体决策还需要考虑群体共识度,即群体决策的目标是在尽可能高的共识程度下寻求能够获得群体最大利益的决策. 而强化学习技术解决问题的过程与多阶段群体的共识达成过程有很多相似之处:1) 都有一个固定的长远目标(达成共识),只有达到了目标才会获得最终的成功,但是在到达目标之前并不确切知道目标的具体状态;2) 每个决策者在给出决策时都仅需要考虑其当前所处的环境状态(包括其他决策者的)和当前策略知识;3) 都强调决策者与环境的交互作用,决策者每作出一个决策都会获得环境的反馈(奖励或者惩罚),决策者下一步的策略依赖于当前反馈和将来所获得奖励. 因此,本文结合强化学习技术来完成多阶段群体决策过程. 在理想状态下,群体决策的目标是在满足群体利益最大的情况下同时满足群体共识最大;同时证明,利用改进的Q-学习算法的多阶段群体决策模型所获得的全局最优决策,同时也是具有群体共识度最大的策略.

本文首先介绍如何利用强化学习技术建立多阶段群体决策问题模型,并改进Q-学习算法以寻求多阶段群体决策问题的最优策略;然后,证明基于改进的Q-学习算法的多阶段群体决策得到的群体最优策略同时也是群体共识度最高的策略;最后,通过计算实例论证模型的可行性及正确性.

1 强化学习的基本概念

强化学习的主要思想是决策个体通过与环境交互,采用试错方法对某一环境状态下的不同动作的效用函数(奖励函数)进行估计,选取那些获得奖励值最大的动作来不断获得策略的改进[10-11],如图1所示.

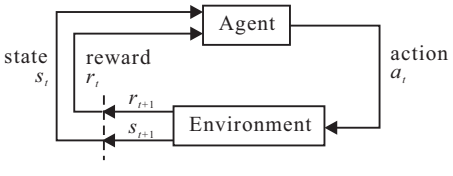

图1 强化学习框架

强化学习不同于监督学习,强化学习没有事先提供的训练样例可以利用,其监督信息只来源于最终决策结束时获得的环境奖赏值. 在多数的决策问题中,每一阶段甚至每一步决策都可能影响最终的决策结果,即最终的奖赏值受到现在所采取一个或多个动作的影响,其目的在于求解一个能使最终奖赏最大的全

局最优策略,这与多阶段决策问题的目的是一致的.

强化学习任务通常是用马尔科夫决策过程(Markov decision process, MDP)来描述:一个MDP对应于一个5元组$E=\langle S,A,P,R,\gamma\rangle$. 其中:$S$为状态空间,每个状态$s\in S$表示决策者感知到的环境的描述;$A$为动作空间,$a\in A$为当前状态下决策者所能采取的动作;$P:S\times A\times S\to[0,1]$指定了状态转移概率,表示若某个动作$a\in A$作用在当前状态$s$上,则将以转移概率$p$使得环境从当前状态转移到另一个状态;$R:S\times A$为奖赏函数;$\gamma$为折扣因子. 决策者通过在环境中不断地执行动作,得到环境反馈,最终不断尝试从而学习到一个"策略"π,根据这个策略可知在状态s下所要执行的动作$a=\pi(s)$. 策略有两种表达方法:一种是将策略表示为函数$\pi:S\to A$,确定性策略常用这种方式表示;另一种是概率表示$\pi:S\times A\to R$,随机性策略常用这种方式表示,$\pi(s,a)$为状态s下选择动作a的概率,其中$\sum\limits_a\pi(s,a)=1$[12].

强化学习经典的学习算法之一是Q-学习算法,它采用状态-动作值函数$Q(s,a)$为估值函数,在每次迭代时考虑整个动作空间的每个动作,并保证学习过程的最终收敛[13]. Q-学习算法[14-18]的基本迭代公式如下:

$$Q(s,a)=$$
$$Q(s,a)+\alpha(r+\gamma\max_a Q(s',a)-Q(s,a)). \tag{1}$$

其中:α为学习率,r为由状态s转移到s'后带来的奖赏值. 由式(1)可知,最优策略即在当前状态s下采用使得Q值最大的动作a.

2 基于强化学习技术的多阶段群体决策模型

设有m个决策者$D_1,D_2,\cdots,D_m$,其权重分别为$(\lambda_1,\lambda_2,\cdots,\lambda_m)$,参与一个包含$T$个阶段的多阶段群体决策问题. 为叙述简洁,将此多阶段群体决策问题描述为一个5元组

$$E=\langle S,A,P,R,\gamma\rangle.$$

其中:$S=\{S_1,S_2,\cdots,S_T\}$为不同阶段的状态空间,$S_i=\{s_i^1,s_i^2,\cdots,s_i^{k_i}\}$,$s_i^k$为决策阶段$i$的第$k$个状态,$i=1,2,\cdots,T$;$A=\{A_1,A_2,\cdots,A_T\}$为动作空间,$A_i\in A$为对应第$i$阶段的状态集$S_i$的备选方案,$A_i=\{a_i^1,a_i^2,\cdots,a_i^{n_i}\}$;$P=\{P_1,P_2,\cdots,P_T\}$,$P_i$为第$i$个阶段的状态转移矩阵,$i=1,2,\cdots,T$;$R$为在这一阶段的某状态采取动作$a$后到达下一阶段的某个状态时,决策者对于此行为的满意值.

令$V_j^\pi(s_{i-1})$表示决策者j从状态$s_{i-1}(i=1,2,\cdots,T)$开始,遵循被评估策略$\pi(s)$,经过T个阶段带来的累积满意值,则有

$$V_j^\pi(s_{i-1})=\sum_{a_i\in A_i}\pi(s_{i-1},a_i)\sum_{s_i\in S_i}P_{s_{i-1}\to s_i}^{a_i}(R_{s_{i-1}\to s_i}^{a_i}+\gamma V(s_i)), \tag{2}$$

$$Q_j^\pi(s_{i-1},a_i)=\sum_{s_i\in S_i}P_{s_{i-1}\to s_i}^{a_i}(R_{s_{i-1}\to s_i}^{a_i}+\gamma V(s_i)). \tag{3}$$

$\boldsymbol{J}^\pi$表示m个决策者对于策略$\pi(s)$的经过T个阶段的群体满意值,有

$$\boldsymbol{J}^\pi(s)=\sum_{j=1}^m\lambda_jV_j^\pi(s), \tag{4}$$

$$\boldsymbol{J}^\pi(s,a)=\sum_{j=1}^m\lambda_jQ_j^\pi(s,a). \tag{5}$$

最后使得群体满意值最大的策略即为多阶段群体决策的最优策略$\pi(s_{i-1})=\arg\max_{a_i}J(s_{i-1},a_i)$.

由此,Q-学习算法的迭代公式在多阶段群体决策中转化为

$$Q_j^\pi(s_{i-1},a_i)=Q_j^\pi(s_{i-1},a_i)+\alpha(R_{s_{i-1}\to s_i}^{a_i}+\gamma\max_a Q(s_i,a)-Q_j^\pi(s_{i-1},a_i)), \tag{6}$$

$$\boldsymbol{J}^\pi(s,a)=\sum_{j=1}^m\lambda_jQ_j^\pi(s_{i-1},a_i). \tag{7}$$

具体算法可描述如下.

算法1　多阶段群体决策的Q-学习算法.

输入:环境E、动作空间A、初始状态s_0、阶段数T、决策者权重$W=(\lambda_1,\lambda_2,\cdots,\lambda_m)$、奖赏折扣$\gamma$、更新步长$\alpha$、决策者个数$m$;

输出:最优群体策略.

Step 1: $Q(S,A)=0,\pi(s_i,a)=\dfrac{1}{|A(s_i)|}$.

Step 2: for $i=1,2,\cdots,T$

for $j=1,2,\cdots,m$ do

根据式(2)和(3)计算

$$Q_j^\pi(s_{i-1},a_i)=Q_j^\pi(s_{i-1},a_i)+\alpha(R_{s_{i-1}\to s_i}^{a_i}+\gamma\max_a Q(s_i,a)-Q_j^\pi(s_{i-1},a_i))$$

end for

end for

Step 3: 根据式(6)计算$\boldsymbol{J}^\pi(s,a)=\sum\limits_{j=1}^m\lambda_jQ_j^\pi(s_{i-1},a_i)$.

Step 4: 寻找$\pi(s_{i-1})=\arg\max_{a_i}J(s_{i-1},a_i)$.

Step 5: 输出群体最优策略π.

从上述算法可以分析得到该算法的时间复杂度为$O(m\times T)$.

群体共识度是一个用于表示群体意见一致性程度的参数,群体的意见越一致表明群体共识度越大.

群体决策问题中通常用于度量群体共识的方法可以分为两种[19]:一种是计算群体满意值与各个决策者满意值之间的距离;另一种是计算各个决策者之间的距离. 本文基于第一种方法度量群体共识.

定义1 给定一个多阶段的群体决策策略π,假设其群体满意值为$\boldsymbol{J}^{\pi}$,参与决策的m个决策者对于该策略的满意值分别为 $Q_1^{\pi},Q_2^{\pi},\cdots,Q_m^{\pi}$,则群体共识度可利用群体决策与个体决策之间的距离定义,即

$$\boldsymbol{C}^{\pi}=\frac{1}{\sqrt{(\boldsymbol{J}^{\pi}-Q_1^{\pi})^2+\cdots+(\boldsymbol{J}^{\pi}-Q_m^{\pi})^2}}. \tag{8}$$

显然,群体共识度$\boldsymbol{C}^{\pi}$越大,表示决策群体对该策略π的满意度越高;反之,群体共识度$\boldsymbol{C}^{\pi}$越小,决策群体对该策略π的满意度越低.

定理1 基于多阶段群体决策的Q-学习算法得到的多阶段群体最优决策$\boldsymbol{\pi}^*(s)$具有最大的群体共识度$\boldsymbol{C}^{\pi}$.

证明 记$\boldsymbol{\pi}^*(s)$为多阶段群体决策问题的最优策略,其群体满意度为$\boldsymbol{J}^{\pi*}$,则由$Q_j^{\pi}(s_{i-1},a_i)=Q_j^{\pi}(s_{i-1},a_i)+\alpha(R_{s_{i-1}\to s_i}^{a_i}+\gamma\max\limits_{a}Q(s_i,a)-Q_j^{\pi}(s_{i-1},a_i))$以及上述方程的收敛性[13]可得群体最优策略为

$$\boldsymbol{\pi}^*(s)=\arg\max_{a\in A}\boldsymbol{J}^{\pi}(s,a).$$

由于群体对于该策略的满意度值为

$$\boldsymbol{J}^{\pi*}(s,a)=\max_{\pi}\sum_{j=1}^{m}\lambda_jQ_j^{\pi}(s,a),$$

从而有

$$\boldsymbol{J}^{\pi*}(s,a)=\max_{\pi}\sum_{j=1}^{m}\lambda_jQ_j^{\pi}(s,a)=$$

$$\sum_{j=1}^{m}\lambda_j\max_{\pi}Q_j^{\pi}(s,a)=\sum_{j=1}^{m}\lambda_jQ_j^{\pi*}(s,a).$$

其中:权重$\lambda_j\geqslant 0$,个体满意度$Q_i^{\pi}\geqslant 0\,(i=1,2,\cdots,m)$.

由此可知,能够使得$\boldsymbol{J}^{\pi}(s,a)$取得最大值的策略$\pi$必然能够使得$Q_1^{\pi}(s,a),Q_2^{\pi}(s,a),\cdots,Q_m^{\pi}(s,a)$都取最大值,即群体最优策略与各决策个体的最优策略是完全一致的,从而表明基于多阶段群体决策的Q-学习算法所得到的多阶段群体最优决策$\boldsymbol{\pi}^*(s)$具有最大的群体共识度$\boldsymbol{C}^{\pi}$. □

3 计算实例

假设某公司要投资一个项目,该项目评估过程分为市场调研、商业计划、产品试验、市场出售4个阶段. 该公司主要是由5个决策者组成的一个决策群体共同参与这个具有4个决策阶段的决策过程,因为决策群体在专业认知、公司地位等级等方面存在差异,所以对该决策过程的权重不同. 设决策者的权重分别为(0.1, 0.1, 0.4, 0.1, 0.3),每个决策者对于每个阶段下各状态的满意值按照十分制进行打分(表1),满分10分,最低分1分(为叙述方便,分数为0表示该阶段没有此状态),令奖赏折扣$\gamma=0.9$,更新步长$\alpha=0.01$,状态转移概率$P=\{P_1,P_2,\cdots,P_T\}$. 不失一般性,假设初始状态的转移服从均匀随机分布,即$P_{s_{i-1}\to s_i}^{a_i}=\dfrac{1}{k_i}$,$k_i$为第$i$个阶段下的状态总数(通常,转移概率可以通过强化学习过程学习得到).

表1 决策者的满意值

s	a	s'	D_1	D_2	D_3	D_4	D_5
		s_1^1	4	1	3	7	3
	a_1^1	s_1^2	8	4	5	6	6
		s_1^3	6	9	1	4	7
		s_1^1	1	6	3	3	4
s_0	a_1^2	s_1^2	5	7	4	7	5
		s_1^3	4	1	6	8	9
		s_1^1	7	5	1	5	4
	a_1^3	s_1^2	3	1	7	7	6
		s_1^3	8	7	4	6	2
		s_2^1	9	9	5	3	1
	a_2^1	s_2^2	2	7	9	1	6
s_1^1		s_2^1	5	5	3	2	7
	a_2^1	s_2^2	6	6	4	5	4
		s_2^1	4	7	5	3	6
	a_2^1	s_2^2	8	5	3	7	3
s_1^2		s_2^1	2	6	1	6	7
	a_2^1	s_2^2	6	8	4	1	6
		s_2^1	1	4	5	3	3
	a_2^1	s_2^2	4	6	3	7	9
s_1^3		s_2^1	2	1	4	4	5
	a_2^1	s_2^2	7	4	5	3	7
		s_3^1	9	5	4	7	4
	a_3^1	s_3^2	2	3	2	9	1
		s_3^3	5	6	5	6	6
		s_3^1	6	6	7	8	3
s_2^1	a_3^2	s_3^2	8	5	5	6	4
		s_3^3	3	2	6	3	9
		s_3^1	6	6	2	1	5
	a_3^3	s_3^2	7	4	4	8	7
		s_3^3	8	1	9	4	4
		s_3^1	2	6	4	7	7
	a_3^1	s_3^2	4	3	6	5	5
		s_3^3	5	1	7	9	6
		s_3^1	3	5	7	3	3
s_2^2	a_3^2	s_3^2	7	9	5	5	6
		s_3^3	3	2	3	6	4
		s_3^1	6	4	2	2	6
	a_3^3	s_3^2	4	8	5	5	7
		s_3^3	9	3	4	3	3
	a_4^1	s_4	6	7	8	4	8
s_3^1	a_4^2	s_4	7	5	5	8	3
	a_4^1	s_4	8	3	7	6	5
s_3^2	a_4^2	s_4	2	7	5	4	4
	a_4^1	s_4	4	9	4	8	3
s_3^3	a_4^2	s_4	3	5	6	5	2

先初始化Q矩阵为0,然后开始实验,在初始状态s_0下按照$\pi(s_0^m,a)=\dfrac{1}{|A(s_0^m)|}$($m$为决策者个数)的概率选取备选方案,根据多阶段群体策略Q-学习基本算法更新Q矩阵,由Q矩阵可以计算出J矩阵,同时状态从初始状态s_o以1/3的概率转移到下一阶段的某一状态,再根据多阶段群体决策Q-学习算法更新J矩阵,同理,更新各个阶段的J矩阵,直到J矩阵

收敛,最终得到收敛矩阵J表达形式如下:

$$J=\begin{bmatrix} J(s_0,a_1^1) & J(s_0,a_1^2) & \dots & J(s_0,a_4^2)\\ J(s_1^1,a_1^1) & J(s_1^1,a_1^2) & \dots & J(s_1^1,a_4^2)\\ \vdots & \vdots & \ddots & \vdots\\ J(s_3^3,a_1^1) & J(s_3^3,a_1^2) & \dots & J(s_3^3,a_4^2)\end{bmatrix}.$$

由上述算法可计算出如下结果:

$$J=\begin{bmatrix} 16.38 & 16.85 & 16.26 & 0 & 0\\ 0 & 0 & 0 & 13.32 & 13.00\\ 0 & 0 & 0 & 13.28 & 13.07\\ 0 & 0 & 0 & 13.25 & 13.00\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}\rightarrow$$

$$\leftarrow\begin{bmatrix} 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 9.36 & 9.07 & 6.60 & 0 & 0\\ 9.33 & 9.04 & 6.60 & 0 & 0\\ 0 & 0 & 0 & 4.37 & 4.99\\ 0 & 0 & 0 & 4.42 & 4.95\\ 0 & 0 & 0 & 4.42 & 4.91 \end{bmatrix},$$

即多阶段群体最优策略为

$$\pi=(a_1^2,a_2^1,a_3^1,a_4^2).$$

其中:a_1^2,a_2^1,a_3^1,a_4^2分别对应于每阶段群体满意值最大时采取的方案. 第1到第4阶段最大群体满意值所对应的方案选择,即群体最优策略与文献[2]中的结果相比,不仅达到了Pareto最优,而且能够获得决策结果的群体最大满意度.

根据定义1中的群体共识度,利用数据(表2)进一步验证由多阶段群体策略Q-学习算法得到的群体策略具有最高的群体共识度.

由于在多阶段群体决策中,每一阶段下选择备选方案后所到达的状态未知,在对每个阶段寻找备选方案时,可对每阶段备选方案求群体共识度期望值. 具体计算如下.

第1阶段:

$$C(s_0,a_1^1)=0.07,\ C(s_0,a_1^2)=0.10,\ C(s_0,a_1^3)=0.09.$$

第2阶段:

$$C(s_1,a_2^1)=\frac{1}{3}C(s_1^1,a_2^1)+\frac{1}{3}C(s_1^2,a_2^1)+\frac{1}{3}C(s_1^3,a_2^1)=0.14,$$

$$C(s_1,a_2^2)=\frac{1}{3}C(s_1^1,a_2^2)+\frac{1}{3}C(s_1^2,a_2^2)+\frac{1}{3}C(s_1^3,a_2^2)=0.11.$$

第3阶段:

$$C(s_2,a_3^1)=\frac{1}{2}C(s_2^1,a_3^1)+\frac{1}{2}C(s_2^2,a_3^1)=0.24,$$

$$C(s_2,a_3^2)=\frac{1}{2}C(s_2^1,a_3^2)+\frac{1}{2}C(s_2^2,a_3^2)=0.17,$$

$$C(s_2,a_3^3)=\frac{1}{2}C(s_2^1,a_3^3)+\frac{1}{2}C(s_2^2,a_3^3)=0.22.$$

第4阶段:

$$C(s_3,a_4^1)=\frac{1}{3}C(s_3^1,a_4^1)+\frac{1}{3}C(s_3^2,a_4^1)+\frac{1}{3}C(s_3^3,a_4^1)=0.18,$$

$$C(s_3,a_4^2)=\frac{1}{3}C(s_3^1,a_4^2)+\frac{1}{3}C(s_3^2,a_4^2)+\frac{1}{3}C(s_3^3,a_4^2)=0.22.$$

表 2　多阶段群体决策值和个体决策值

s	a	J	Q_1	Q_2	Q_3	Q_4	Q_5	$\boldsymbol{C}^{\pi}$
s_0	a_1^1	16.38	19.77	18.58	14.16	18.91	20.00	0.07
	a_1^2	16.85	17.10	18.67	15.4	19.18	20.61	0.10
	a_1^3	16.26	18.42	18.28	15.16	19.22	18.59	0.09
s_1^1	a_2^1	13.32	15.28	15.54	12.41	13.58	15.12	0.14
	a_2^2	13.00	13.72	14.45	11.22	14.63	16.12	0.11
s_1^2	a_2^1	13.28	15.38	15.58	12.44	13.47	15.03	0.14
	a_2^2	13.07	13.61	14.44	11.25	14.62	16.19	0.12
s_1^3	a_2^1	13.25	15.36	15.55	12.42	13.48	15.07	0.14
	a_2^2	13.00	13.63	14.55	11.21	14.72	16.19	0.11
s_2^1	a_3^1	9.36	10.91	10.49	8.66	9.42	10.28	0.23
	a_3^2	9.07	9.24	9.39	7.38	10.58	11.34	0.17
	a_3^3	6.60	7.76	6.38	5.80	8.39	7.35	0.21
s_2^2	a_3^1	9.33	10.94	10.33	8.67	9.36	10.22	0.24
	a_3^2	9.04	9.24	9.45	7.40	10.61	11.39	0.16
	a_3^3	6.60	7.70	6.42	5.99	8.36	7.31	0.23
s_3^1	a_4^1	4.37	6.04	4.68	2.97	5.64	5.28	0.18
	a_4^2	4.99	3.33	4.28	4.34	5.99	5.99	0.20
s_3^2	a_4^1	4.42	6.01	4.87	3.02	5.64	5.27	0.18
	a_4^2	4.95	3.33	4.77	4.40	6.03	5.93	0.23
s_3^3	a_4^1	4.42	6.00	4.70	3.00	5.66	5.29	0.19
	a_4^2	4.91	3036	4.77	4.32	5.99	5.97	0.23

由此可验证,策略$\pi = (a_1^2, a_2^1, a_3^1, a_4^2)$既是多阶段群体最优策略,又是群体共识度最高的策略. 因此多阶段群体最优策略同时满足群体共识度最高.

4 结 论

本文研究了一种在不确定状态空间下求解多阶段群体决策的最优决策的方法. 利用强化学习在环境中可以通过试错-学习的特点,通过改进强化学习的经典Q-学习算法,提出了一种在不确定状态环境下求解多阶段群体决策的最优决策的算法. 理论验证表明,该算法得到的群体最优解不仅是群体满意度最大的解,同时也是具有最高群体共识度的最优决策. 最后,通过一个计算实例验证了算法的合理性及可行性. 本文的主要创新点在于将强化学习的Q-学习算法应用到多阶段群体决策问题中,提出了一种能够在不确定的环境下通过不断学习环境知识来进行最优决策的算法,同时在一定程度上丰富了群体决策问题的内容.

参考文献(References)

[1] Hwang C L. Group decision making under multiple criteria, methods and applications[M]. Berlin: Springer -Verlay, 1987: 311-317.

[2] 彭怡, 胡杨. 多阶段群体决策的Pareto最优策略[J]. 四川大学学报: 自然科学版, 2007, 44(3): 482-484.
(Peng Y, Hu Y. The Pareto optimization policies of multi-stage group decision making[J]. J of Sichuan University: Natural Science Edition, 2007, 44(3): 482-484.)

[3] 郝晶晶, 朱建军, 刘远. 双重信息下多阶段异质群体决策模型计算[J]. 系统工程，2016, 34(5): 129-134.
(Hao J J, Zhu J J, Liu Y. Model and algorithm for multi-stage group decision-making concerning different decision groups and dual information[J]. Systems Engineering, 2016, 34(5): 129-134.)

[4] Zhang Y J. Optimal algorithm of multistage group decision-making under discrete determinate status[J]. J of China Three Gorges University, 2013, 35(2): 104-107.

[5] Lu Z P, Lu C Y. Fast aggregated model for multi-stage group decision-making probelm based on preference distance method[C]. Int Conf on Management and Service Science. Wuhan: IEEE, 2011: 1-4.

[6] 张娜，方志耕，朱建军. 基于Orness测度约束的多阶段灰色局势群体决策模型[J]. 控制与决策, 2015, 30(7): 1227-1232.
(Zhang N, Fang Z G, Zhu J J. Multi-stage grey situation group decision-making model based on Orness[J]. Control and Decision, 2015, 30(7): 1227-1232.)

[7] Luo D, Li Y W. Multi-stage and multi-attribute risk group decision-making method based on grey information[C]. Workshop on Grey System Theory and its Applications. Beijing: China Center of Advanced Science and Technology, 2014: 305-310.

[8] 周声海. 基于模糊偏好关系的多目标多阶段冲突型复杂大群体决策方法研究[D]. 长沙: 中南大学商学院, 2013.
(Zhou S H. Research on the large group decision making method of multi-objective and multi-stage within conflict based on fuzzy preference relation[D]. Changsha: Business School of Central South University, 2013.)

[9] 马跃如, 王雄. 多时段条件下动态物流联盟伙伴选择的模糊语言群体决策模型[J]. 系统工程, 2008, 26(6): 32-36.
(Ma Y R, Wang X. Partner selection model based on fuzzy language group decision-making method for dynamic logistics alliance under multiple time periods[J]. Systems Engineering, 2008, 26(6): 32-36.)

[10] Kaelbling L P. Reinforcement learning: A survey[J]. J of Artificial Intelligence Research, 1996, 4(1): 237-285.

[11] Zheng S L, Han J H, Luo X F, et al. Research on cooperation and reinforcement learning in multi-agent systems[J]. Pattern Recognition & Artificial Intelligence, 2002, 15(4): 453-456.

[12] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2015: 371-390.
(Zhou Z H. Machine learning [M]. Beijing: Tsinghua University Press, 2015: 371-390.)

[13] Tom M Mitchell. 机器学习: 计算机科学丛书[M]. 机械工业出版社, 2014: 270-271.
(Tom M Mitchell. Machine learning[M]. China Machine Press, 2014: 270-271.)

[14] Hao J, Huang D, Cai Y, et al. The dynamics of reinforcement social learning in networked cooperative multiagent systems[J]. Engineering Applications of Artificial Intelligence, 2017, 58: 111-122.

[15] Zhan Y S, Ammar H B, Taylor M E. Scalable lifelong reinforcement learning[J]. Pattern Recognition, 2017, 72: 407-418.

[16] Foerster J, Nardelli N, Farquhar G, et al. Stabilising experience replay for deep multi-agent reinforcement learning[C]. Proc of the 34th Int Conf on Machine Learning. Sydney: Proc of Machine Learning Research, 2017: 1146-1155.

[17] Mnih V, Kavukcuoglu K, Silver D, et al. Playing atari with deep reinforcement learning[EB/OL]. (2013-12-19)[2018-01-16]. https://www.cs.toronto.edu/vmnih/docs/dqn.pdf.

[18] Mnih V, Kavukcuoglu K, Silver D, et al, Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[19] Palomares I, Martinez L. A semi-supervised multiagent system model to support consensus-reaching processes[J]. IEEE Trans on Fuzzy Systems, 2014, 22(4): 762-777.

作者简介

张峰(1976−), 女, 副教授, 博士, 从事机器学习、智能决策等研究, E-mail: amyfzhang@yahoo.com;

刘凌云(1993−), 女, 硕士生, 从事机器学习的研究, E-mail: 1058738029@qq.com;

郭欣欣(1992−), 女, 硕士生, 从事机器学习的研究, E-mail: 1871467157@qq.com.

(责任编辑：闫 妍)