计算机科学与探索
Journal of Frontiers of Computer Science and Technology
1673-9418/2024/18(02)-0378-09
doi: 10.3778/j.issn.1673-9418.2210090

开放科学(OSID)

融合强化学习的三支治略选择及其有效性分析

刘晓雪[1],姜春茂[2+]
1. 哈尔滨师范大学 计算机科学与信息工程学院,哈尔滨 150025
2. 福建理工大学 计算机科学与数学学院,福州 350118
+ 通信作者 E-mail: hsdrose@126.com

摘　要:三支决策的“分、治、效”(TAO)模型包括构建三分、施加策略、结果评估三个部分。目前,关于结果评估的研究旨在衡量策略施加后结果的前后变化,还无法预测施加哪个策略能达到最大效果。为了解决这一问题,对TAO模型的“治”和“效”进行了研究,提出一种基于强化学习的三支改变模型的策略选择与有效性预测的方法。首先将改变三支决策TAO模型中的改变三分状态和策略分别作为强化学习中的状态和动作,并将每次施加策略得到新的改变三分状态的过程看作一个周期,利用累积前景理论计算每个周期产生的奖励,将智能体与环境的交互过程用马尔可夫决策过程来表示;其次设置一个目标奖励,将各个周期的累计奖励达到目标奖励时的状态作为马尔可夫决策过程的终止状态;然后用Q-learning算法迭代出一个最短周期内达到目标奖励的策略序列,同时利用该策略序列预测当前改变三分状态的未来效用。最后使用一个实例体现出该方法实用性和有效性。

关键词:三支决策;改变三支决策;强化学习;策略选择;效用度量
文献标志码:A　　**中图分类号:**TP18

Strategy Selection and Outcome Evaluation of Three-Way Decisions Based on Reinforcement Learning

LIU Xiaoxue[1], JIANG Chunmao[2+]
1. School of Computer Science and Information Engineering, Harbin Normal University, Harbin 150025, China
2. School of Computer Science and Mathematics, Fujian University of Technology, Fuzhou 350118, China

Abstract: The trisecting-acting-outcome (TAO) model of three-way decision (3WD) consists of three steps: trisect a whole, design action strategies, and outcome analysis and measurement. Currently, research on outcome evaluation aims to measure the pre- and post-change in outcomes following the implementation of strategies, and it is still unable to predict which strategy will achieve the maximum effect. To narrow down this gap, this paper focuses on the “acting” and “outcome” of the TAO model and introduces a method for strategy selection and outcome prediction for the change-based three-way decision based on Q-learning in reinforcement learning. Firstly, the approach is to treat the altered tri-partition and the acting in the change-based three-way decision TAO model as states and actions in reinforcement learning, respectively, and to consider the process of obtaining a newly altered tri-partition each time under the acting of action or strategy as a cycle. The reward generated by each cycle is calculated using cumulative prospect theory, and the interaction process between the agent and the environment is represented by a Markov decision process. Secondly, a target reward is set, and the state when the cumulative reward of each cycle reaches

基金项目:黑龙江省自然科学基金(LH2020F031)。
This work was supported by the Natural Science Foundation of Heilongjiang Province (LH2020F031).
收稿日期:2022-10-24　　**修回日期:**2023-04-03

the target reward is taken as the termination state of the Markov decision process. Then a Q-learning algorithm is used to iterate a set of actions that achieve the target reward in the shortest cycle and then the action set is used to predict the future utility of the change-based three-way decision. Finally, an example is employed to illustrate the applicability and effectiveness of the method.

Key words: three-way decision; change-based three-way decision; reinforcement learning; strategy selection; outcome evaluation

三支决策[1-3]是一种基于人类认知的决策模式，它最初由姚一豫教授提出，用于处理不确定、不完备的信息，其核心思想是将一个整体区域划分成三个两两不相交的区域或部分，对于不同的部分分别采取行动或者策略。从狭义的角度来看，三支决策可以描述为接受（acceptance）、延迟（deferment）和拒绝（rejection）；从广义的角度来看，三支决策是基于“三”的思维方式、问题求解方法和信息处理模式[4]。近年来，有关三支决策的研究发展迅速，例如三支属性约简[5-11]、序贯三支决策[12-17]、三支聚类[18-21]、移动三支决策[22-24]、三支概念分析[25-26]等[27-29]。

在基于分治模型（trisecting-and-acting）的基础上，姚一豫教授进一步提出了“效（outcome）”这一要素，从而形成了“分、治、效”结合的三支决策TAO模型[30]。目前，关于三支决策TAO模型的研究主要集中在“分”，对于“治”和“效”的研究是有限的。在关于“治”的研究中，文献[31]针对不利区域中的对象制定移动策略；文献[32]根据用户的偏好和三分区的结构特征，使用条件熵和交叉熵度量比例来选择最佳的动作和策略；文献[33]提出了一个带有多阶段区域转化的三支决策模型。改变三支决策TAO模型是对三支决策TAO模型的扩展，它的思想是使对象从原来的三分区改变到更有效的三分区，描述了一种基于改变状态的三支决策评估模型。目前改变三支决策的研究主要集中在效用度量方面[34-36]。以往三支决策更多关注的是对整体分而治之，而改变三支决策是根据前后状态的改变来衡量其有效性。

上述文献虽然在TAO模型的“治”和“效”上取得了一定的成果，但是它们都需要在实际环境中实施治略后才能评估效用，这增加了成本和风险。而且它们也没有给出一种有效的方法来快速找到最优的治略序列。

为了解决上述问题，本文将强化学习和三支决策相结合，利用强化学习的试错机制来预测能获得最大效用的治略。同时，本文采用强化学习中的Q-learning算法来迭代求解一个最短周期内达到目标奖励的最优治略序列，并利用该治略序列对未来效用进行预测。将强化学习和三支决策进行交叉研究，充分利用了强化学习的学习机制，提高了三支决策的实用性和普适性，使其能够适应更多的复杂环境和任务，同时增强了三支决策的学习能力和自适应能力，使其能够从数据中不断更新和改进。

本文首先回顾了改变三支决策与强化学习的相关基础知识，然后介绍了一种基于Q-learning算法的策略选择方法，并利用累积前景理论计算改变三分状态转移产生的奖励，接着通过一个网店运营的实例验证了本文方法的有效性和实用性，最后总结全文。

1 相关工作

1.1 改变三支决策TAO模型

三支决策的TAO模型包括“分（trisecting）”“治（acting）”“效（outcome）”三个部分，而改变三支决策TAO模型同样聚焦于此，不过又略有区别。改变三支决策TAO模型的“分（trisecting）”是根据对象的改变状态将整体粒化成独立的三个部分；“治（acting）”是通过施加一些策略，来改变全集中对象的质量，从而提高改变模型的有效性；“效（outcome）”是分和治的有效性评估以及指导反馈。

1.1.1 TAO模型的“分”

改变视角的TAO模型是一种特殊类型的三支决策TAO模型，它根据对象改变状态的不同将整体分成三个区域：理想区（R^+）、不确定区（R^0）和不理想区（R^-）。然后对这三个区域施加相应的策略，将发生理想改变的对象划分到R^+区域，发生不确定改变的对象划分到R^0区域，发生不理想改变的对象划分到R^-区域。为了描述这个想法，给出一些相关的定义：

定义1[34] 假设整体U是一个有限非空对象集合，S是一个有限的状态集。基于改变状态集S，全集U被划分到R^+、R^0和R^-三个区域。三分区域$\pi_c=\{R^+,R^0,R^-\}$表示全集U的改变三分区，并且满足以下三个属性：

（1）$R^+ \bigcup R^0 \bigcup R^- = U$

（2）$R^+ \bigcap R^0 = \varnothing$，$R^+ \bigcap R^- = \varnothing$，$R^0 \bigcap R^- = \varnothing$

（3）$R^+ > R^0 > R^-$

治略作用于对象后，对象会发生不同程度的变化。本文通过设定一对改变阈值和一个改变评价函数，来对受治略影响的对象进行划分，这里假设改变评价函数为 $e(U|S)$，表示一个对象 $x \in U$ 的改变量的改变评价值，有如下定义：

定义2 给定对象 $x \in U$ 与改变评价值 $e(x|s)$，并引入一对阈值 (α,β)，令 $\alpha > \beta$，将一个非空有限对象集 U 进行三分，当改变量值大于等于 α 记为 R^+，当改变量的值小于等于 β 记为 R^-，否则，记为 R^0。则非空有限对象集 U 可以被划分成如下三个区域：

$$R^+(U)=\{x \in U|e(x|s) \geq \alpha\}$$

$$R^0(U)=\{x \in U|\beta < e(x|s) < \alpha\}$$

$$R^-(U)=\{x \in U|e(x|s) \leq \beta\}$$

1.1.2 TAO模型的“治”

为了得到更优的三分状态，决策者通常会制定一些有价值的策略来引导对象发生期望的改变。对于同一个改变三分状态，可能有多种可选的策略。不同的策略在不同的时间和空间条件下施加，都会对对象的效用产生影响，从而导致不同的改变三分状态。

在相同的三分标准下，每对全集 U 施加一次策略，就会产生新的改变三分状态，本文将改变三支决策与强化学习相结合，把改变三分状态作为强化学习中的状态，把TAO模型中的治略作为强化学习中的动作，把每次施加策略后得到新的改变三分状态的过程视为一个周期。每个周期的改变三分状态是独立的，它的分布仅与前一个周期的改变三分状态分布有关，与之前各个周期的改变三分状态以及其他因素无关，改变三分状态的动态改变过程具有“无后效性”，因此每个周期中改变三分状态的变化过程是一个马尔可夫过程。

1.1.3 TAO模型的“效”

TAO模型的“效”取决于三分和治略的适当匹配。在先前的研究中，大多是通过判断策略的施加是否有效来评估效用，然后选择最优的策略。假设 U 是一个有限非空集合，π_c 表示对集合 U 产生的三分区治略后得到的改变三分区域，假设 $Q:\pi_c \rightarrow R$ 表示改变三分区域的度量函数，给定改变三分区域 $\pi_c=\{R^+,R^0,R^-\}$，可以有如下的基本度量框架[34]：

$$Q(\pi_c)=\omega_1 Q(R^+)+\omega_2 Q(R^0)+\omega_3 Q(R^-) \tag{1}$$

这里，$Q(R^\chi)$ 是区域 R^χ 的效用，$\chi=\{+,0,-\}$；ω_i 表示不同区域的权重，其中 $\omega_1+\omega_2+\omega_3=1$。

当对改变三分区域 π_c 分别施加相应的策略后，会得到一个新的改变三分区域 π_c'，由此可以利用下式表示基本的度量框架：

$$Q(\pi_c'|\pi_c)=Q(\pi_c')-Q(\pi_c) \tag{2}$$

这里 $Q(\pi_c'|\pi_c)$ 表示施加策略后效用的改变，$Q(\pi_c')$ 和 $Q(\pi_c)$ 分别表示三分区 π_c' 和三分区 π_c 的效用。

1.2 强化学习

强化学习（reinforcement learning，RL）[37]是一种机器学习的方法，它强调如何让智能体在环境中通过自己的行为和反馈的奖励来学习最优的策略。假设环境是马尔可夫型的，则顺序型强化学习问题可以通过马尔可夫决策过程建模。

改变三分状态随着策略的变化是符合马尔可夫性质的，即下一个改变三分状态只依赖于当前的状态和策略，而与之前的历史无关。马尔可夫决策过程是一个由四个元素构成的元组 $<\pi_c,A,\tilde{P},R>$。其中 π_c 是一个包含所有改变三分状态的有限集合；A 是一个包含所有策略的有限集合；$\tilde{P}$ 为状态转移概率 $\tilde{P}(\pi_c'|\pi_c,A)$，表示在状态 π_c 和策略 A 的条件下，下一个状态为 π_c' 的概率；奖励函数 $R(\pi_c,A,\pi_c')$ 表示在状态 π_c 和策略 A 的条件下，转移到状态 π_c' 时获得的奖励。由于马尔可夫决策过程（Markov decision processes，MDP）可以很好地捕捉智能体与环境之间的动态关系，通常使用它来描述强化学习任务。

Q-learning[38]算法是一种强化学习的方法，它可以在马尔可夫决策过程（MDP）中寻找最优策略。它的基本思想是通过不断地探索和学习来更新一个Q表，这个Q表记录了每个状态动作对的价值函数，也就是执行某个动作后能获得的期望回报。Q-learning算法的目标是让Q表收敛到最优的价值函数，也就是在任何状态下都能选择最优的动作。

在强化学习与改变三支决策结合起来研究的过程中，可以将改变三分状态变化产生的效用作为奖励，用来计算更新Q函数，并根据Q函数来选择能够最大化累积奖励的策略。由于在不断的策略下，改变三分状态会不断发生变化，为了简化问题和提高效率，本文提出设定一个目标奖励值，当改变三分状态在多次连续策略下产生的奖励累计达到目标奖励值时，就认为一次探索的结束。按照这样的思路进行多次探索，直到能够稳定找到最短的策略路径。

2 基于强化学习的策略选择及其效用度量

本章利用马尔可夫决策过程来建模改变三分状态通过不断优化策略选择来最大化效用的过程。首先利用二次规划法求解状态转移概率，然后根据累积前景理论计算发生状态改变产生的奖励，最后利用Q-learning算法来寻找每个周期中能够获得最大奖励的策略。

图1表示利用Q-learning算法来更新Q-表格的过程，对于t周期中的改变三分状态π_c^t，在Q-表格中利用ε-贪心策略选择策略A对其施加，在策略A的作用下，可以得到t+1周期中更有效的改变三分状态π_c^{t+1}，从而产生奖励r_t，利用产生的奖励来更新Q-表格。按照这个过程，进行多次训练，直到Q-表格中的Q值达到稳定。下面是对改变三分状态$\{R^+,R^0,R^-\}$进行一系列策略训练的具体思路：

(1)状态转移概率矩阵。计算每个策略施加所带来的转移概率矩阵，根据转移概率矩阵，得到下一个周期中的改变三分状态。

(2)利用累积前景理论计算每一次施加策略产生新的改变三分状态的效用值，并将其作为强化学习中的奖励。

(3)策略选择。利用ε-贪心策略在Q-表格中选出

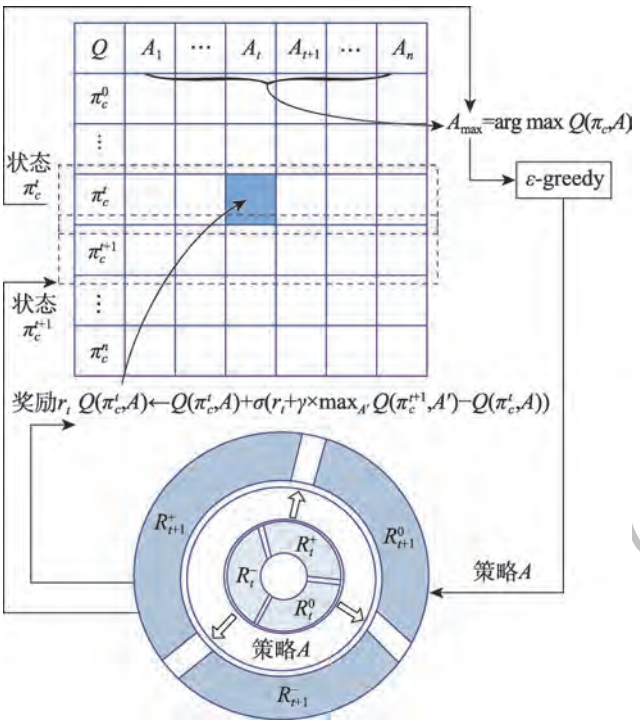

图1 利用Q-learning算法进行三支改变的策略选择

Fig.1 Q-learning algorithm for strategy selection using change-based three-way decision

每一个周期中对应的三分状态应该施加的策略。

(4)设置训练的终止状态。在策略训练的过程中设定一个目标奖励值，当训练获得的累计奖励达到目标奖励时，训练停止。

(5)进行多次训练，直到Q-表格中的Q值达到稳定，然后选择每个周期中Q值最大的策略。

2.1 状态及状态转移概率矩阵

在关于改变三支决策的马尔可夫决策过程中，将随着周期变化的改变三分状态序列作为马尔可夫决策过程中的状态集合。施加不同的策略，会产生不同的状态转移概率矩阵，根据状态转移概率矩阵来判断下一个周期中改变三分状态的状态分布。本节利用二次规划法来求解不同周期之间的状态转移概率矩阵。

设一轮模拟中存在n个周期，每个周期策略施加后产生一个改变三分状态，初始周期中改变三分状态分布的向量为：

$$\boldsymbol{\pi}_c(0)=\left[|R_0^+|,|R_0^0|,|R_0^-|\right]$$

其中，$|R_0^\chi|,\chi=\{+,0,-\}$表示初始周期中改变三分区域的分布的情况。

经过k个周期，对象处于改变三分区R_k^χ的概率估计值为$|R_k^\chi|$，则k步转移后状态分布向量为：

$$\boldsymbol{\pi}_c(k)=[|R_k^+|,|R_k^0|,|R_k^-|]$$

令在马尔可夫预测模型中，状态转移概率矩阵的估计值为$\tilde{\boldsymbol{P}}$：

$$\tilde{\boldsymbol{P}}=\begin{bmatrix}\tilde{p}_{++} & \tilde{p}_{+0} & \tilde{p}_{+-}\\ \tilde{p}_{0+} & \tilde{p}_{00} & \tilde{p}_{0-}\\ \tilde{p}_{-+} & \tilde{p}_{-0} & \tilde{p}_{--}\end{bmatrix}$$

于是$\boldsymbol{\pi}_c(k)$的估计值为$\tilde{\boldsymbol{\pi}}_c(k)$：

$$\tilde{\boldsymbol{\pi}}_c(k)=\boldsymbol{\pi}_c(k-1)\tilde{\boldsymbol{P}}=[|R_{k-1}^+|,|R_{k-1}^0|,|R_{k-1}^-|]\cdot\tilde{\boldsymbol{P}}=\boldsymbol{\pi}_c(0)\cdot\tilde{\boldsymbol{P}}^k=[|R_0^+|,|R_0^0|,|R_0^-|]\cdot\begin{bmatrix}\tilde{p}_{++} & \tilde{p}_{+0} & \tilde{p}_{+-}\\ \tilde{p}_{0+} & \tilde{p}_{00} & \tilde{p}_{0-}\\ \tilde{p}_{-+} & \tilde{p}_{-0} & \tilde{p}_{--}\end{bmatrix}^k,k=1,2,\cdots,n \tag{3}$$

令$C_\chi(k)$表示经过k步转移后，改变三分区域R^χ的概率估计值$|\tilde{R}_k^\chi|$与实际值$|R_k^\chi|$的误差：

$$C_\chi(k)=|R_k^\chi|-|\tilde{R}_k^\chi|=|R_k^\chi|-\sum_{j=\{-,0,+\}}^{|j|}|R_{k-1}^j|\tilde{P}_{\chi j} \tag{4}$$

其中，$j=\{+,0,-\}$，$\chi=\{+,0,-\}$，且$+>0>-$。

那么n个周期，改变三分区域R^χ的误差的平方和E_χ：

$$E_\chi=\sum_{k=1}^{n}(C_\chi(k))^2 \tag{5}$$

则在各个周期改变三分区域中对象的转移过程中的误差平方和为 D：

$$D=\sum_{\chi=\{-,0,+\}}^{|\chi|}E_{\chi}=\sum_{\chi=\{-,0,+\}}^{|\chi|}\sum_{k=1}^{n}(C_{\chi}(k))^2=\sum_{\chi=\{-,0,+\}}^{|\chi|}\sum_{k=1}^{n}\left(|R_k^{\chi}|-\sum_{j=\{-,0,+\}}^{|j|}|R_{k-1}^{j}|\tilde{P}_{\chi j}\right)^2 \tag{6}$$

根据最小二乘法的思想将约束条件引入模型中，建立一个以各个状态下各个阶段转移过程中误差的平方和 D 最小为目标函数，以行和条件和非负条件为约束条件的求解马尔可夫状态转移概率矩阵的优化模型，具体如下：

$$\begin{cases}\min D=\sum_{\chi=\{-,0,+\}}^{|\chi|}\sum_{k=1}^{n}\left(|R_k^{\chi}|-\sum_{j=\{-,0,+\}}^{|j|}|R_{k-1}^{j}|\tilde{P}_{\chi j}\right)^2\\ \text{s.t.}\sum_{j=\{-,0,+\}}^{|j|}P_{\chi j}=1\\ P_{\chi j}\geqslant 0\end{cases} \tag{7}$$

二次规划的求解方法很多，如Lemke方法、有效集法、线性转换法等。这里将其转换成线性规划模型进行求解，降低模型的求解难度。

2.2 策略与策略选择

对于每个周期中的改变三分状态，存在多个可施加的策略，将它们看作马尔可夫决策过程中的动作集合。本文的目标是选择一种策略方式，去最大化改变三分状态从最初状态到终止状态所能获得的累计奖励。为此，本节采用Q-learning算法，通过不断的探索和学习来更新一个记录着每个状态动作对的价值函数的Q-表格。

在进行策略选择时，首先构建列标签为改变三分状态 π_c^i，行标签为可施加的策略 A_i 的Q-表格。在初始周期中，将Q-表格中的数值全部初始化为0。针对Q-表格中对应周期的改变三分状态，采用 ε-贪心策略，选择准备施加的策略，具体来讲是当策略对应的动作值函数 $Q(\pi_c,A)$ 大于 ε 或者 $Q(\pi_c,A)$ 值全为0时，就随机选取一个策略来施行，反之则选取对应周期中 $Q(\pi_c,A)$ 最大值对应的策略。

Q-表格中的 $Q(\pi_c,A)$ 值是不断更新的，那么对于 $Q(\pi_c,A)$ 值的更新方式具体如下：

假设对周期 i 的改变三分状态 π_c^i 施加策略 A_m，则其真实的 $Q(\pi_c^i,A_m)_{\text{Re}}$ 计算公式为：

$$Q(\pi_c^i,A_m)_{\text{Re}}=r_i+\gamma\times\max_{A'}Q(\pi_c^{i+1},A') \tag{8}$$

其中，r_i 表示在第 i 周期下施加策略 A_m 产生的改变三分状态 π_c^i 的奖励值，γ 表示折扣因子，$\max_{A'}Q(\pi_c^{i+1},A')$ 表示Q表格中第 $i+1$ 周期的改变三分状态 π_c^{i+1} 对应的最大的 $Q(\pi_c^{i+1},A')$。

而在Q-表格中改变三分状态 π_c^i 施加策略 A_m 对应的位置上 $Q(\pi_c^i,A_m)$ 值是之前训练后更新完成的价值函数，将其作为估计值 $Q(\pi_c^i,A_m)_{\text{Es}}$。

$CQ(\pi_c^i,A_m)$ 表示估计值和真实值之间的差距：

$$CQ(\pi_c^i,A_m)=Q(\pi_c^i,A_m)_{\text{Re}}-Q(\pi_c^i,A_m)_{\text{Es}} \tag{9}$$

根据上述公式更新Q-表格中改变三分状态 π_c^i 与施加策略 A_m 对应位置的 $Q(\pi_c^i,A_m)$ 值：

$$Q(\pi_c^i,A_m)\leftarrow Q(\pi_c^i,A_m)_{\text{Es}}+\sigma CQ(\pi_c^i,A_m) \tag{10}$$

其中，σ 为学习率。

对于Q-表格的更新，人工的计算是无法完成的，因此可以利用Q-learning算法来进行计算，算法的思想如下所示：

算法1 Q-learning算法更新Q-表格

输入：初始的改变三分状态 π_c^0，训练次数 E，目标奖励 R，累计奖励 CR，学习率 σ，折扣因子 γ。

输出：更新完成的Q-表格。

1. Initialize Q-table
2. Create a tri-partition dynamic change environment with transfer probability and reward functions
3. for i in range(E):
4. Initialize π_c
5. while $CR\leqslant R$:
6. Select A from Q-table using ε-greedy
7. Taking A gets π_c' and r from the environment
8. $Q(\pi_c,a)\leftarrow Q(\pi_c,a)_{\text{Es}}+\sigma CQ(\pi_c,a)$
9. $CR\leftarrow CR+r$
10. $\pi_c\leftarrow\pi_c'$
11. end while
12. end for
13. return Q-table

根据上述算法，可以得到一个更新完成的Q-表格，为了能够在最短时间内达到目标奖励，可以在更新完成的Q-表格中选择每个改变三分状态下Q值最大的策略，组成一个最佳的策略序列。该策略序列是通过过往的经验学习到的能够获得最大效用的策略，因此能够为之后的策略选择提供帮助。

2.3 奖励与未来效用

在改变三支决策的马尔可夫决策过程中，每次改变三分状态的变化都会产生一个奖励，利用该奖励来更新Q-表格。本节利用累积前景理论的思想，对每个周期产生的奖励进行计算。

累积前景理论表示的是决策者在存在风险和不确定性下的决策行为和风险态度，它的主要思想可

以概括为三个方面[39]：首先，结果被视为相对于参考点的收益或损失，而不是财富的最终状态；其次，决策者对收益厌恶风险，对损失寻求风险，对损失比收益更敏感；最后，面对极小概率事件，决策者倾向于放大它的影响，面对极大概率事件，决策者倾向于缩小它的影响。累积前景理论指出决策者更喜欢具有最大累积前景值的决策选项。借助值函数 $v(x_h)$ 和决策权重函数 ψ_h，累积前景值函数 υ 表示为：

$$\upsilon=\sum_{h=-m}^{n}\psi_h v(x_h) \tag{11}$$

对于第 k 周期中的改变三分状态 $\{R_k^+,R_k^0,R_k^-\}$，对其施加策略 A 就会使全集 U 中的对象发生区域变化，产生第 $k+1$ 周期的改变三分状态 $\{R_{k+1}^+,R_{k+1}^0,R_{k+1}^-\}$。由于策略 A 的施加，R_k^+ 区域中的对象可能分别发生理想的、不确定的或者不理想的改变。同理，R_k^0、R_k^- 区域中的对象也可能有这样的改变趋势。假设 x_k 表示在 R_k^χ 区域的对象改变到 R_{k+1}^χ 产生的效用，决策者需设置一个参考值 $\bar{x}$，若 R_k^χ 区域中的对象在策略 A 的作用下，发生了理想的和不确定的改变，则 $x_k\geqslant\bar{x}$，表示改变产生了收益，否则就产生了损失。根据实验验证，Tversky 和 Kahneman[40]提供了值函数 $v(x_h)$ 的详细公式：

$$v(x_h)=\begin{cases}(x_h-\bar{x})^{\mu}, x_h\geqslant\bar{x}\\ -\theta(\bar{x}-x_h)^{v}, x_h<\bar{x}\end{cases} \tag{12}$$

这里令 $\mu=v=0.88,\theta=2.25$ [39]。根据上述公式可以看出值函数是单调递增函数。

将改变效用值 x_k 代入求值函数 $v(x_h)$ 的公式，可以得到在策略 A 的作用下，由第 k 周期的改变三分状态得到更有效的第 k+1 周期的改变三分状态各区域改变的值函数具体如表 1。

表 1　对象区域改变产生的值函数

Table 1　Value functions for object region changes

π_c^k	π_c^{k+1}		
	R_{k+1}^+	R_{k+1}^0	R_{k+1}^-
R_k^+	v_{++}^k	v_{+0}^k	v_{+-}^k
R_k^0	v_{0+}^k	v_{00}^k	v_{0-}^k
R_k^-	v_{-+}^k	v_{-0}^k	v_{--}^k

在累积前景理论中，决策者对于收益与损失的权重函数计算方式是不一样的，因此考虑了概率的两种非线性变换，分别给出在收益和损失的情况下的权重函数计算公式为：

$$w^+(p_h)=\frac{p_h^\sigma}{(p_h^\sigma+(1-p_h)^\sigma)^{1/\sigma}};\ w^-(p_h)=\frac{p_h^\delta}{(p_h^\delta+(1-p_h)^\delta)^{1/\delta}} \tag{13}$$

p_h 是 x_{ij}^k 对应的概率，这里令 $\sigma=0.61,\delta=0.69$ [39]。

在连续概率变换中，需要对累积概率进行加权，并且对各个改变区域中对象的改变效用按升序排序，分别针对收益或损失，每个效用的决策权重函数 ψ_h 可以表示为：

$$\psi_h=\begin{cases}w^+(p_h+\cdots+p_n)-w^+(p_{h+1}+\cdots+p_n), h\geqslant 0\\ w^-(p_{-m}+\cdots+p_h)-w^-(p_{-m}+\cdots+p_{h-1}), h<0\end{cases} \tag{14}$$

通过上述对于值函数的计算可以知道值函数是单调递增函数，因此可以将改变产生的值函数进行排序，以获得每种改变相关的权重函数。通过对比各个值函数的大小，可以得到所有情况下的决策权重函数，表示如下：

$$\psi_i(\Pr(R_{k+1}^\chi[\![x]\!]))=\begin{cases}w^+(\Pr(R_{k+1}^+[\![x]\!])), \chi=+\\ w^+(\Pr(R_{k+1}^+[\![x]\!])+\Pr(R_{k+1}^0[\![x]\!]))-w^+(\Pr(R_{k+1}^+[\![x]\!])), \chi=0\\ w^-(\Pr(R_{k+1}^-[\![x]\!])), \chi=-\end{cases} \tag{15}$$

其中，$i=\{+,0,-\},\chi=\{+,0,-\},\Pr(R_{k+1}^\chi[\![x]\!])$ 表示等价类 $[x]$ 属于区域 R_{k+1}^χ 的概率。

因此根据累积前景理论公式可以算出第 k 周期的改变三分区中各个区域的累积前景值 υ：

$$\upsilon(R_k^+[\![x]\!])=\psi_+(\Pr(R_{k+1}^+[\![x]\!]))v_{++}^k+\psi_+(\Pr(R_{k+1}^0[\![x]\!]))v_{+0}^k+\psi_+(\Pr(R_{k+1}^-[\![x]\!]))v_{+-}^k \tag{16}$$

$$\upsilon(R_k^0[\![x]\!])=\psi_0(\Pr(R_{k+1}^+[\![x]\!]))v_{0+}^k+\psi_0(\Pr(R_{k+1}^0[\![x]\!]))v_{00}^k+\psi_0(\Pr(R_{k+1}^-[\![x]\!]))v_{0-}^k \tag{17}$$

$$\upsilon(R_k^-[\![x]\!])=\psi_-(\Pr(R_{k+1}^+[\![x]\!]))v_{-+}^k+\psi_-(\Pr(R_{k+1}^0[\![x]\!]))v_{-0}^k+\psi_-(\Pr(R_{k+1}^-[\![x]\!]))v_{--}^k \tag{18}$$

因此第 k 周期施加策略 A 得到的改变三分状态 π_c^{k+1} 产生的奖励为：

$$r_k=\upsilon(R_k^+[\![x]\!])+\upsilon(R_k^0[\![x]\!])+\upsilon(R_k^-[\![x]\!]) \tag{19}$$

根据上述奖励的计算方式，可以得到每次训练过程中每个周期产生的奖励，进而预估改变三支决策的未来效用。利用第 2.2 节中的 Q-learning 算法来更新 Q-表格，并根据 Q-表格找出一条最短路径达到目标奖励的策略序列。结合状态转移概率矩阵，可以预测出一条未来的改变三分状态序列，并用累积前景理论计算每个周期的奖励。在这个过程中，用下式来计算改变三支决策的未来效用 $G(\pi_c)$：

$$G(\pi_c)=r_1+\gamma r_2+\cdots+\gamma^{n-1}r_n \tag{20}$$

其中，r_i 表示在模拟的最短周期中，第 i 周期获得的奖励。

由于该方法在预测未来效用的过程中使用累积前景理论计算每个周期产生的奖励，更能真实地反映出决策者的行为偏好。利用该方法计算出的未来

效用不仅能够在一系列策略前就能够预估结果，而且预测的结果相对来说也更加准确。

3 应用实例

本章使用一个店铺运营的例子来阐述本文的思想，这里是选择出能使店铺在有限的周期内最快达到目标奖励的策略方式，并预测当前状态下，店铺的未来收益。

网店运营者为了提升销量，会定期给店内产品做各种平台活动。根据活动的情况，店铺内产品的销量会发生不同程度的变化。根据产品销量的变化情况，若某一产品的销量增加量超过10，就将其划分为爆款；若某一产品的销量减少10，就将其划分为潜力款；否则就是促销款，分别用 R^+、R^- 和 R^0 表示。将店铺每策划一次活动使店铺内产品销量变化的过程看作一个周期，一般每个周期持续3~5天。

目前平台上有三个引流活动 a_1、a_2、a_3，对于店铺中的产品，分别针对 R^+、R^0、R^- 三个不同的区域施加合适的引流活动，将每个周期对三个区域施加的活动组合起来作为一次策略，店铺中产品销量随着所做策略的不同呈现出不同的改变情况。根据往期的经验，现设置三组策略 $A_1=\{a_1,a_3,a_2\}$、$A_2=\{a_3,a_1,a_2\}$、$A_3=\{a_2,a_2,a_1\}$。

3.1 一次状态转移过程

现从店铺中选择76件对活动敏感的产品作为采样数据。在往期的店铺数据中，分别截取策略 A_1、A_2、A_3 的数据各20个周期。根据此数据分别得到策略 A_1、A_2、A_3 初始周期中的爆款、潜力款、促销款三档产品的数量分布为[21,30,25]、[15,36,25]、[28,21,27]。在第20个周期后，三档产品的产品数量分布分别为[46,14,16]、[36,25,15]、[42,17,17]。利用2.1节中的方法，可以得出店铺产品在策略 A_1、A_2、A_3 的作用下，发生状态转移的转移概率约分别为：

$$\tilde{\boldsymbol{P}}_{A_1}=\begin{bmatrix}0.603\,7 & 0.145\,2 & 0.251\,1\\ 0.565\,5 & 0.223\,1 & 0.211\,4\\ 0.657\,6 & 0.231\,0 & 0.111\,4\end{bmatrix}$$

$$\tilde{\boldsymbol{P}}_{A_2}=\begin{bmatrix}0.366\,2 & 0.524\,2 & 0.109\,6\\ 0.514\,5 & 0.121\,3 & 0.364\,2\\ 0.657\,6 & 0.231\,0 & 0.111\,4\end{bmatrix}$$

$$\tilde{\boldsymbol{P}}_{A_3}=\begin{bmatrix}0.576\,8 & 0.214\,2 & 0.209\,0\\ 0.548\,4 & 0.220\,3 & 0.231\,3\\ 0.514\,6 & 0.214\,1 & 0.271\,3\end{bmatrix}$$

假设在第 k 周期中，对产品数量分布为[17,28,31]的销量变化三分区域施加策略 A_1，就会产生新的产品三分区域[46,16,14]，并产生奖励。接下来利用2.3节中的方法计算奖励，首先根据产品分布变化的情况，为各区域中设置的参数如表2所示。

表2 各区域参数

Table 2 Regional parameters

π_c	$\chi=+$	$\chi=0$	$\chi=-$
R_k^{χ}	17	28	31
R_{k+1}^{χ}	46	16	14
x_k	$1\,000\times\lvert R_k^i\cap R_{k+1}^{\chi}\rvert$	$500\times\lvert R_k^i\cap R_{k+1}^{\chi}\rvert$	$-200\times\lvert R_k^i\cap R_{k+1}^{\chi}\rvert$

表2中 $i=\{+,0,-\}$，令 $\bar{x}=0.1$。根据式(12)，可以得到状态改变产生的值函数，如表3所示。

表3 改变产生的值函数

Table 3 Value function generated by change

π_c^k	π_c^{k+1}		
	R_{k+1}^+	R_{k+1}^0	R_{k+1}^-
R_k^+	3 323.81	441.97	−1 045.62
R_k^0	4 901.08	1 115.21	−1 323.80
R_k^-	6 138.44	1 268.34	−904.53

然后根据式(15)，计算各区域改变的权重，结果如表4所示。

表4 各区域改变的权重值

Table 4 Weight values generated by change in each region

π_c^k	π_c^{k+1}		
	R_{k+1}^+	R_{k+1}^0	R_{k+1}^-
R_k^+	0.476 0	0.091 5	0.294 3
R_k^0	0.455 0	0.143 0	0.265 6
R_k^-	0.507 2	0.221 1	0.181 5

根据上述值函数和权重，结合式(19)，可以求得第 k 周期的收益值为：

$$r^k=v(R_k^+[x])+v(R_k^0[x])+v(R_k^-[x])\approx$$
$$1\,314.85+2\,037.87+3\,229.67=6\,582.39$$

3.2 多次模拟结果分析

现对当前产品销量的改变三分区[12,26,38]进行未来效用预测，假设店铺的目标收益为50 000，令 ε-贪心策略中的 ε 等于0.8，学习率 σ 等于0.1，折扣因子 γ 等于0.8。根据2.2节中的Q-learning的算法来更新Q-表格。

在训练达到300次左右的时候，达到目标收益所需要的周期数稳定到了9个周期。图2展示了多次训练达到目标收益所需要的周期数的变化趋势。从图中可以看出，随着训练次数的增加，达到目标收益所需要的周期呈现下降的趋势。训练结束后，将Q-表格导出如表5所示。

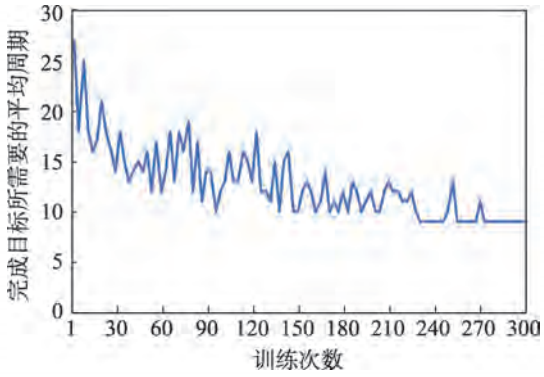

图2 多次模拟的平均周期变化情况

Fig.2 Average cycle variation of multiple simulations

表5 更新完成的Q-表格

Table 5 Updated Q-table

π_c^k	策略		
	A_1	A_2	A_3
π_c^0	25 126.68	26 025.18	22 150.23
π_c^1	24 679.51	22 530.43	23 636.89
π_c^2	20 943.68	21 041.82	22 999.41
π_c^3	21 205.15	22 118.56	20 760.78
π_c^4	19 472.07	19 443.12	20 958.95
π_c^5	19 299.01	17 351.48	15 284.67
π_c^6	16 219.77	14 032.83	14 973.53
π_c^7	12 378.92	11 154.01	12 427.71
π_c^8	8 682.98	6 834.03	7 683.58

根据表5可以看出，在9个周期内依次施加策略$\{A_2,A_1,A_3,A_2,A_3,A_1,A_1,A_3,A_1\}$可以最快达到目标收益，这为店铺之后的策略选择提供了帮助，按照该策略方案，可以提高店铺运营的效率和收益。

在最后一次模拟中，第0~8个周期产生的奖励分别约为6 988.68、6 279.39、5 442.3、5 733.48、5 464.21、6 451.19、6 386.86、6 394.74、6 393.75，因此可以计算当前改变三分状态π_c的未来效用如下：

$$G(\pi)=6\,988.68+0.8\times6\,279.39+0.8^2\times5\,442.3+0.8^3\times5\,733.48+0.8^4\times5\,464.21+0.8^5\times6\,451.19+0.8^6\times6\,386.86+0.8^7\times6\,394.74+0.8^8\times6\,393.75\approx26\,870.91$$

上述改变三分状态的未来效用是基于最优策略预测出的最大收益，它对评估店铺当前的运营状况具有重要参考价值。本文方法能够更好地帮助店铺进行策略选择与效用评估，从而指导运营者更好地经营和管理店铺，实现收益最大化，提升市场份额和竞争力，因而具有极其重要的现实意义。

4 结束语

本文针对改变三支决策TAO模型的“治”和“效”进行了讨论，将强化学习引入改变三支决策的策略选择与效用度量中，利用累积前景理论计算每个周期状态转移产生的奖励，并将其作为反馈信号，利用Q-learning算法更新Q-表格，根据更新完成的Q-表格选出最短周期内达到目标奖励的治略序列，并预测未来效用。

本文将强化学习引入到三支决策模型中，提高了决策效率和质量，并为其他类型决策模型提供了新颖而有效的解决方案。然而，本文也存在一些不足之处，如未考虑改变三分区划分的阈值随环境变化的问题，以及未描述根据具体问题和环境来调整Q-learning算法参数的方法。未来将在这些方面进一步研究。

参考文献：

[1] YAO Y Y. Three-way decision: an interpretation of rules in rough set theory[C]//LNCS 5589: Proceedings of the 4th International Conference on Rough Sets and Knowledge Technology, Gold Coast, Jul 14-16, 2009. Berlin, Heidelberg: Springer, 2009: 642-649.

[2] YAO Y Y. Three-way decisions with probabilistic rough sets [J]. Information Sciences, 2010, 180(3): 341-353.

[3] YAO Y Y. An outline of a theory of three-way decisions[C]// LNCS 7413: Proceedings of the 8th International Conference on Rough Sets and Current Trends in Computing, Chengdu, Aug 17-20, 2012. Berlin, Heidelberg: Springer, 2012: 1-17.

[4] YAO Y Y. The geometry of three-way decision[J]. Applied Intelligence, 2021, 51(9): 6298-6325.

[5] FANG Y, MIN F. Cost-sensitive approximate attribute reduction with three-way decisions[J]. International Journal of Approximate Reasoning, 2019, 104: 148-165.

[6] JIA X Y, LIAO W H, TANG Z M, et al. Minimum cost attribute reduction in decision-theoretic rough set models[J]. Information Sciences, 2013, 219: 151-167.

[7] MIN F, ZHU W. Attribute reduction of data with error ranges and test costs[J]. Information Sciences, 2012, 211: 48-67.

[8] JIA X Y, SHANG L, ZHOU B, et al. Generalized attribute reduce in rough set theory[J]. Knowledge-Based Systems, 2016, 91: 204-218.

[9] MA X A, ZHAO X R. Cost-sensitive three-way class-specific attribute reduction[J]. International Journal of Approximate Reasoning, 2019, 105: 153-174.

[10] QIAN J, DANG C Y, YUE X D, et al. Attribute reduction for sequential three-way decisions under dynamic granulation [J]. International Journal of Approximate Reasoning, 2017, 85: 196-216.

[11] ZHANG X Y, YANG J L, TANG L Y. Three-way class-specific attribute reducts from the information viewpoint[J]. Information Sciences, 2020, 507: 840-872.

[12] YAO Y Y. Granular computing and sequential three-way de-

cisions[C]//LNCS 8171: Proceedings of the 8th International Conference on Rough Sets and Knowledge Technology, Halifax, Oct 11-14, 2013. Berlin, Heidelberg: Springer, 2013: 16-27.
[13] ZHANG Q H, PANG G H, WANG G Y. A novel sequential three-way decisions model based on penalty function[J]. Knowledge-Based Systems, 2020, 192: 105350.
[14] YANG X, LI T, FUJITA H, et al. A sequential three-way approach to multi-class decision[J]. International Journal of Approximate Reasoning, 2019, 104: 108-125.
[15] QIAN J, LIU C H, MIAO D Q, et al. Sequential three-way decisions via multi-granularity[J]. Information Sciences, 2020, 507: 606-629.
[16] ZHANG L B, LI H X, ZHOU X Z, et al. Sequential three-way decision based on multi-granular autoencoder features [J]. Information Sciences, 2020, 507: 630-643.
[17] JU H R, PEDRYCZ W, LI H X, et al. Sequential three-way classifier with justifiable granularity[J]. Knowledge-Based Systems, 2019, 163: 103-119.
[18] YU H, WANG X C, WANG G Y, et al. An active three-way clustering method via low-rank matrices for multi-view data [J]. Information Sciences, 2020, 507: 823-839.
[19] WANG P X, YAO Y Y. CE3: a three-way clustering method based on mathematical morphology[J]. Knowledge-Based Systems, 2018, 155: 54-65.
[20] YU H, ZHANG C, WANG G Y. A tree-based incremental overlapping clustering method using the three-way decision theory[J]. Knowledge-Based Systems, 2016, 91: 189-203.
[21] AFRIDI M K, AZAM N, YAO J T, et al. A three-way clustering approach for handling missing data using GTRS[J]. International Journal of Approximate Reasoning, 2018, 98: 11-24.
[22] JIANG C M, YAO Y Y. Effectiveness measures in movement-based three-way decisions[J]. Knowledge-Based Systems, 2018, 160: 136-143.
[23] JIANG C M, GUO D D, DUAN Y, et al. Strategy selection under entropy measures in movement-based three-way decision[J]. International Journal of Approximate Reasoning, 2020, 119: 280-291.
[24] JIANG C M, GUO D D, XU R Y. Measuring the outcome of movement-based three-way decision using proportional utility functions[J]. Applied Intelligence, 2021, 51(12): 8598-8612.
[25] QI J J, QIAN T, WEI L. The connections between three-way and classical concept lattices[J]. Knowledge-Based Systems, 2016, 91: 143-151.
[26] WEI L, LIU L, QI J J, et al. Rules acquisition of formal decision contexts based on three-way concept lattices[J]. Information Sciences, 2020, 516: 529-544.
[27] YANG B, LI J H. Complex network analysis of three-way decision researches[J]. International Journal of Machine Learning and Cybernetics, 2020, 11: 973-987.
[28] LIU D, LIANG D C, WANG C C. A novel three-way decision model based on incomplete information system[J]. Knowledge-Based Systems, 2016, 91: 32-45.
[29] LI J H, HUANG C C, QI J J, et al. Three-way cognitive concept learning via multi-granularity[J]. Information Sciences, 2017, 378: 244-263.
[30] YAO Y Y. Three-way decision and granular computing[J]. International Journal of Approximate Reasoning, 2018, 103: 107-123.
[31] GAO C, YAO Y Y. Actionable strategies in three-way decisions[J]. Knowledge-Based Systems, 2017, 133: 141-155.
[32] JIANG C M, ZHAO S B. Action strategy analysis in probabilistic preference movement-based three-way decision[J]. Mathematical Problems in Engineering, 2020. DOI: 10.1155/2020/5436507.
[33] 郭豆豆, 姜春茂. 基于 M-3WD 的多阶段区域转化策略研究[J]. 计算机科学, 2019, 46(10): 279-285.
GUO D D, JIANG C M. Multi-stage regional transformation strategy in movement-based three-way decision model [J]. Computer Science, 2019, 46(10): 279-285.
[34] JIANG C M, GUO D D, SUN L J. Effectiveness measure for TAO model of three-way decisions with interval set[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(6): 11071-11084.
[35] GUO D D, JIANG C M, SHENG R X, et al. a novel outcome evaluation model of three-way decision: a change viewpoint [J]. Information Sciences, 2022, 607: 1089-1110.
[36] JIANG C M, GUO D D, DUAN Y. Measure effectiveness of change-based three-way decision using utility theory[J]. Cognitive Computation, 2022, 14(3): 1009-1018.
[37] KAELBLING L P, LITTMAN M L, MOORE A W. Reinforcement learning: a survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237-285.
[38] WATJINS C J C H, DAYAN P. Q-learning[J]. Machine Learning, 1992, 8(3): 279-292.
[39] WANG T X, LI H X, ZHANG L B, et al. A three-way decision model based on cumulative prospect theory[J]. Information Sciences, 2020, 519: 74-92.
[40] TVERSKY A, KAHNEMAN D. Advances in prospect theory: cumulative representation of uncertainty[J]. Journal of Risk and Uncertainty, 1992, 5(4): 297-323.

刘晓雪(1997—),女,河南开封人,硕士研究生,主要研究方向为三支决策、强化学习、粗糙集理论等。

LIU Xiaoxue, born in 1997, M.S. candidate. Her research interests include three-way decision, reinforcement learning, rough set, etc.

姜春茂(1972—),男,博士,教授,CCF 会员,主要研究方向为三支决策、云计算、粒计算等。

JIANG Chunmao, born in 1972, Ph.D., professor, CCF member. His research interests include three-way decision, cloud computing, granular computing, etc.