文章编号:1005-0523(2020)01-0047-07

基于强化学习的平行航班动态定价

方 园,乐美龙

(南京航空航天大学民航学院,江苏 南京 21100)

摘要:由于平行航班之间的竞争越来越激烈,为提高航空公司收益,对机票销售系统中的航班和旅客分别建模。将航班的动态定价问题建模成马尔可夫博弈过程,对混合类型旅客建立Logit选择模型。利用多Agent的强化学习算法对实例进行求解,结果表明WoLF-PHC算法收敛所需迭代的次数大于Nash-Q算法,但在计算速度上WoLF-PHC算法优势明显,且具有较强的适应能力。此外,航空机票的定价策略与其他易逝品有所不同,整体呈现上升趋势。而旅客环境参数的变化,也会影响定价策略。基于WoLF-PHC算法得到的定价策略对于收益提升具有积极作用。

关键词:平行航班;混合型旅客;动态定价;马尔可夫博弈;强化学习

中图分类号:[U-9] **文献标志码**:A

DOI:10.16749/j.cnki.jecjtu.2020.01.007

随着航运市场的不断扩大,各航空公司之间的竞争越来越激烈。大部分航线会有多家航空公司同时运营,这些具有相同起讫点的航班称为平行航班。价格成为影响其收益的主要因素。航空公司之间的竞争以及供求关系的变化导致航线产品的价值在不断变化,不论是对航空公司还是旅客来说,进行动态定价十分必要。现有机器学习和大数据技术的发展以及航空市场的宽松环境也为动态定价提供了有利的环境。

原有的较多收益管理都从舱位分配的角度进行研究[1-2],还有将舱位控制和定价同时进行的[3],现在逐渐转向从价格的角度进行收益管理。Gallego G等[4]最早研究多产品下的动态定价问题,他们将单产品的研究拓展到多产品的动态定价中,采用强度控制的方法求解多产品的动态规划问题,导出相应的Hamilton-Jacobi-Bellman方程得到渐进最优解。Zhang D等[5]用马尔可夫决策过程对多个相同目的地之间的可替代航班进行动态定价。通过上下限和启发式的方法进行求解,并通过定期更新各时点的状态来改变价格。Akcay Y等[6]建立了一种能捕捉纵向和横向的产品异质性的线性随机效用框架,并在此基础上进行多产品联合动态定价,采用微分方程进行求解。Gallego G等[7]采用连续时间上的随机博弈,他们利用仿射函数逼近的方法求解微分博弈,并获得渐近均衡解。曹海娜[8]和朱志愚等[9]采用MNL(multinomial logit model)选择模型对平行航班进行动态定价。以上这些研究大多采用近似的方法对NP问题进行求解,使得求解结果不够准确。目前有部分研究利用强化学习方法解决动态定价问题。Han W等[10]通过提前建立其他主体的决策模型并预测他们的价格来构建自身的定价模型,使得问题从多主体决策转变成单主体决策,并采用改进的Q学习方法求解。王金田等[11]和陆慧[12]根据消费者的选择行为是否易受折扣的影响,将其分为两大类,采用强化学习对双卖家市场进行动态定价,但对消费者选择行为的考虑较为简单。Rana R等[13-14]利用带资格迹的强化学习方法分析高峰时刻和非高峰时刻的定价区别,并考虑多个具有关联性产品的定价问题。通过为其它关联性产品对需求的影响设置参数,问题仍转变为单主体的强化学习问题。文献[15-17]研究智能电网的最优动态定价问题。通过顾客消费模型构建定价的环境,同时将动态零售定价问题转化为有限离散马尔可夫决策过程(MDP),并采用Q学习求解最优定价策略。本文将在以往研究的基础上增加对旅客类型,以及需求在各时间

收稿日期:2019-05-15

基金项目:江苏省自然科学基金项目(20151479);中央高校基本科研业务费专项资金资助项目(NZ2016109)

作者简介:方园(1996—),女,硕士研究生,研究方向为航空公司收益管理。

通讯作者:乐美龙(1964—),男,教授,博士,研究方向为航空系统优化和航空公司收益管理。

段内的异质性的考虑,并将智能电网相关研究中所采用的单主体动态定价方法拓展至平行航班的多主体动态定价问题中。相比于以往研究,得到的定价策略更加合理精确,且能有效地捕捉不同类型旅客的选择行为,并有效提升航空公司收益。

1 系统模型

本系统中包含两个重要的角色,分别是旅客和航空公司。假设两家航空公司各自运营的一个航班为平行航班,且出发时刻较为接近。两个航班的旅客群体设为 I。由于航线产品属于易逝品,具有一定长度的销售区间。在以往的大部分研究中,通过对销售区间进行细分,使得每个时间段至多只有一名旅客到达。但这种细分方式导致状态非常多,使得实际问题的求解时间大大增加。另外,在航空公司的实际定价中,也不可能每销售出去一个座位就重新定价,这会使得旅客产生负面情绪。因此,本文考虑以天为单位,将销售区间分为 T 个时间段,每个时间段 t 表示一天。通过机票可销售价格的离散化,将其定义为价格集合 A。集合的大小是确定且有限的。

在每个时间段 t,两家航空公司分别确定该航班的票价,旅客则根据剩余舱位数和当前票价确定自身的购票需求。航空公司的目标是实现自身收益最大化。在销售区间结束后,舱位没有剩余价值。假设系统不考虑超售和团队旅客,且两个平行航班的初始舱位数和各个时间段内的可选择价格都已知。

1.1 旅客行为建模

旅客行为具有两个重要特征,分别是到达率和估值。假设旅客的到达率 λ 服从泊松分布。此外,还需考虑旅客在两个竞争航班之间的选择行为。MNL 模型是一个随机效用最大化模型,用来描述旅客在平行航班间的选择。假设每个旅客购买机票 i 后获得的产品效用为:$U_i=u_i+\varepsilon_i$,其中 $u_i=v_i+\eta f_i$。因此,$U_i=v_i+\eta f_i+\varepsilon_i$,$i=1,2$。$v_i$ 表示航班 i 的平均价值,η 表示价格反应系数,f_i 为机票的定价,ε_i 为随机变量, 用来描述不能观测到的效用随机项。假设 ε_i 服从二重指数分布且各变量两两相互独立,将旅客放弃购买航班机票的效用定义为 0,则每个旅客的购买概率为

$$q_i(f)=P\{U_i=\max_{j=0,1,2} U_j\}=\frac{\exp(v_i+\eta f_i)}{1+\sum_{j=1}^{2}\exp(v_i+\eta f_i)} \qquad i=1,2 \tag{1}$$

旅客放弃购买任何一个航班机票的概率为

$$q_i(f)=\frac{1}{1+\sum_{j=1}^{2}\exp(v_i+\eta f_i)} \tag{2}$$

考虑到旅客的策略性行为,短视型旅客在到达的当前时间段,会通过效用模型计算购买概率,若效用值小于 0,则会放弃购买;若大于 0,则选择购买效用高的航线。而策略型旅客不止看到当期的效用,还会和未来预期的效用进行比较,这也就造成了计算的复杂性。但旅客对于未来预期的收益并不能做到完全理性的判断,只是一个自身的经验估计值。本文利用历史该阶段售价的均值计算未来预期的效用。

根据其策略性行为和估值,将所有的旅客分为 4 大类:高估值策略型旅客,高估值短视型旅客,低估值策略型旅客,低估值短视型旅客。C_θ 表示顾客类型,C_{HS},C_{HM},C_{LS},C_{LM} 分别表示以上 4 种类型旅客。ω 为策略型旅客所占的比例,则 $\overline{\varpi}=1-\overline{\omega}$ 为短视型旅客占比。在策略型旅客中,高估值旅客所占的比例为 φ_S,则低估值旅客所占的比例为 $1-\varphi_S$,表示为 $\overline{\varphi_S}$。同理,短视型旅客中高估值旅客所占比例为 φ_m,低估值旅客所占比例为 $\overline{\varphi_m}$。w_θ 表示某种类型旅客在总体旅客中所占的比例,则该 4 种旅客比例分别为 $w_{\theta=HS,HM,LS,LM}=\{\varpi*\varphi_S,\overline{\varpi}*\varphi_m,\varpi*\overline{\varphi_S},\overline{\varpi}*\overline{\varphi_m}\}$。$\beta_\theta$ 表示不同类型旅客的策略程度,其中,短视型旅客不具有策略性,则 $\beta_{HM,LM}=0$。$\beta_{LS},\beta_{HS}>0$ 通过在仿真环境中设置好这些参数以实现旅客行为的仿真。

1.2 航线产品定价方建模

单航班的动态定价问题通常采用马尔可夫决策过程(MDP)建模。将单个航空公司的动态定价问题拓展

至平行航班的定价问题上,构建多主体的随机博弈 Γ。对一个具有两个玩家和多个状态的随机博弈而言,可以看成是MDP 和矩阵博弈的组合,它是将 MDP 拓展至两个 Agent 上,也是将矩阵博弈拓展多个状态上。随机博弈同样具有马尔可夫性质。

航线产品的定价方称为 Agent,分别用(i=1,2)表示。两个航班的座位总数分别用 N_1,N_2 表示。x_i^t 表示航班 i 在时间段 t 的剩余舱位数。随机博弈可用元组表示为$(S,A_1,A_2,r_1,r_2,p,\gamma)$。$S$ 表示状态空间,用当前时间段和两个航班的剩余舱位数表示,$s^t=(t,x_1^t,x_2^t)$。在每个时间段内,有旅客到达并且购票,状态会发生变化,时间段 t 会减 1,两个航班的库存水平的状态也会根据当前时间段售出的舱位数而发生变化。

A 为可选择的行动集合,每个航班的可选择价格集合为 $A_i=(f_1,\cdots,f_n)$,表示每个 Agent i 的行动空间,$a_i\in A_i$。p 为执行行动后的转移概率,代表在当前状态下,两个玩家执行行动后转移至下一个状态的概率,$\sum_{s'\in S}p(s'|s,a_1,a_2)=1$。$r_i$ 为在每个玩家 i 和对手玩家做出决策后的回报值,和当前时段售出的座位数以及当前票价有关。γ 是马尔可夫决策过程中的折扣因子,$\gamma\in[0,1)$。整个马尔可夫博弈可以看作是在给定状态 s 下,两个 Agent 通过博弈独立的选择行动 a_1,a_2 后,Agent i 收到回报值 $r_i(s,a_1,a_2)$,并且会根据联合行动及转移概率转移至下一个状态 s'。对整个随机博弈而言,假设 π_i 为每个 Agent i 在整个决策过程中采用的策略,在此问题下,即决策者在每个状态下选择各个价格的概率。在给定初始状态 s 后,Agent i 想要最大化:

$$V_i(s,\pi_1,\pi_2)=\sum_{t=0}^{T}\gamma^t E(r_i^t|\pi_1,\pi_2,s_0=s) \tag{3}$$

式中:$E(r_i^t|\pi_1,\pi_2,s_0=s)$表示在初始状态 s,策略为 π_1,π_2 时,在 t 时间段的期望收益值。则 $V_i(s,a_1,a_2)$为从 0 开始到结束的总收益。当航班起飞后,不具有剩余价值。因此,最后的收敛条件为:$V_i((x,N,n),\pi_1,\pi_2)=0$ 或 $V_i((x,n,N),\pi_1,\pi_2)=0$;$V_i((0,x_1,x_2),\pi_1,\pi_2)=0$。销售区间结束后,或者当某个航班的舱位全部售出后的收益值为 0。当满足这两个条件之一时,平行航班之间的竞争结束。

每个特定状态的矩阵博弈称为阶段博弈(stage game)。由于两个航班属于不同航空公司,具有竞争关系,在每个状态下的阶段博弈中寻找纳什均衡策略以实现收益最大化。

定理 1　在一个阶段博弈中的纳什均衡可以描述为 n 个均衡策略的元组,使得 $V_i((s,\pi_1^*,\cdots,\pi_i^*,\cdots,\pi_N^*)\geqslant V_i((s,\pi_1^*,\cdots,\pi_i,\cdots,\pi_N^*)$*for all* $\pi_i\in\prod_i$。

用 $V_i^*(s)$表示纳什均衡策略下的状态值函数,$Q^*(s,a_1,a_2)$表示在遵循纳什均衡策略下的行动值函数,则

$$V_i^*(s)=\sum_{a_1,a_2\in A_1\times A_2}Q_i^*(s,a_1,a_2)\pi_1^*(s,a_1)\cdot\pi_2^*(s,a_2) \tag{4}$$

$$Q_i^*(s,a_1,a_2)=\sum_{s'\in S}T(s,a_1,a_2,s')[r_1(s,a_1,a_2,s')+\lambda V_i^*(s')] \tag{5}$$

式中:$\pi_i^*(s,a_2)\in PD(A_i)$是在玩家 i 的纳什均衡策略下在行动 a_i 上的概率分布。$T(s,a_1,a_2,s')=p(s_{k+1}=s'|s_k=s,a_1,a_2)$是在给定状态和联合行动后转移至该状态的概率。

由此,式(5)中的纳什均衡可以重写为

$$\sum_{a_1,a_2\in A_1\times A_2}Q_i^*(s,a_1,a_2)\pi_1^*(s,a_1)\cdot\pi_i^*(s,a_2)\geqslant\sum_{a_1,a_2\in A_1\times A_2}Q_i^*(s,a_1,a_2)\pi_j^*(s,a_1)\cdot\pi_i^*(s,a_2) \tag{6}$$

通过在每个阶段博弈中求解纳什均衡,以此作为在每个状态下各 Agent 能获得的回报值,根据这个回报值进而学习最优定价策略。

2　算法设计

通过对旅客和航空公司定价方分别建模,利用强化学习算法求解马尔可夫博弈。强化学习不需要知道环境的具体模型,只依赖获得的奖励学习最优行动,它是多 Agent 系统的自然选择。将旅客的选择行为作为

强化学习的环境,每个 Agent 通过与环境交互学习最优策略。在单 Agent 的强化学习环境中,环境是相对稳定的。而多 Agent 系统中,环境中还包括其他 Agent 的行动和状态,即其他 Agent 策略的改变也会影响自身最优策略,因此环境是动态多变的,这对算法的收敛性会带来影响。此外,在单主体的强化学习中,需要存储动作状态 Q 值。而多主体环境中,随着主体的增加,状态空间也增大,联合动作空间呈指数型增长。因此,多智能体系统的维度非常大,计算也变得更加复杂。本文通过对时间段和剩余座位数都做了相应的处理以减少状态数。

2.1 环境设置

环境设置的目标是创建一个虚拟的旅客人群,用来模拟在竞争市场中对市场策略的反应,作为强化学习的学习环境。由于旅客自身的异质性及其购票时的策略行为,将旅客分为 4 种类型。对每种类型的旅客,对其设置到达概率、策略程度和离散选择模型的参数。具体步骤如下:

1) 根据该时段内各类型旅客的到达率模拟旅客的到达数;

2) 根据旅客的离散选择模型和策略程度确定各类型旅客在该价格下的选择概率,若概率大于 0,则选择购买概率高的航班;若概率小于 0,则放弃购买。

2.2 Nash-Q 算法

在多主体环境中, 对 Q 学习算法进行拓展, 将最优 Q 值定义为在 Nash 均衡中收到的 Q 值, 表示为 Nash-Q 值。计算纳什均衡需要已知自身和对方的收益值和状态值,因此需要维护多个 Q 值表。在每一次阶段博弈中利用 Lemke-Howson 算法计算纳什均衡解:$(Q_1^t(s_{t+1},\cdot),\cdots,Q_M^t(s_{t+1},\cdot))$,从而计算出在各状态下的各 Agent 均衡收益值 $NashQ_i^t(s)$。用这个值对每个 Agent 的 Q 值进行更新。Q 值的更新规则为

$$Q_i^{t+1}(s,a_1,\cdots,a_M)=(1-a_t)Q_i^t(s,a_1,\cdots,a_M)+a_t[r_i^t+\gamma NashQ_i^t(s')] \tag{7}$$

其中 a_t 是学习率,M 为 Agent 的个数。通过这些 Q 值的迭代计算,最后收敛至一个稳定值,从而获得最优定价策略。据此,Nash-Q 算法的流程如下所示:

步骤一:初始化时间段 t,初始状态 s_0,对每个 Agent 设置索引 i;

步骤二:对所有的 $s\in S$,以及 $a_i\in A_i$,初始化 $Q_i^t(s_t,a_1,\cdots,a_M)=0$;

步骤三:与旅客环境进行交互,通过 Lemke-Howson 算法获得纳什均衡解,确定所有玩家的行动 a_i^t 后,从而计算收益值 $r_t^1,\cdots,r_t^M$ 并确定下一个状态 $s_{t+1}=s'$,对每个 i,利用公式(7)更新 Q 值;

步骤四:让 $t=t+1$,若 t 为最终状态对应的时间段,则结束该循环;否则,返回至步骤三。

其中各个阶段的状态 S,做出了相应的简化。状态中 x_i^t,不用真实的剩余座位数来表示,而是将一个区间范围内的剩余座位数投射到一个值上以表示当前剩余座位数的状态。这大大减少计算过程所需的存储空间,也加快了求解速度。

2.3 WoLF-PHC 算法

策略梯度爬升(PHC,policy hill climbing)算法的本质是在混合策略空间中表现出梯度爬升。PHC 方法不需要知道玩家最近执行行动的信息,以及对手当前策略的信息,这可以减少算法所需的存储空间并且符合实际的竞争环境。非 Agent 选择最高值行动的概率会以学习率 $\delta\in(0,1]$增加,因此策略在不断提升。PHC 算法能保证在算法中学习的 Agent 是理性的,即如果其他玩家的策略收敛至稳定的策略时,那么自身的学习策略也会收敛至对其他玩家策略的最佳反应策略。

WoLF-PHC 算法是 PHC 算法的拓展[18]。它的关键点为:① 两个学习率;② 采用平均策略来近似均衡策略以确定输赢。WoLF 准则用来修正学习率。算法有两个不同的学习率:δ_w 表示赢时的学习率,δ_l 表示输时的学习率,δ_l 大于 δ_w。当 Agent 输时,学习的要比赢的时候快,这使得当 Agent 学习得比期望糟糕时,能对其他 Agent 策略的变化适应得更快;当学习得比期望好时,要学得更谨慎。这也给其他 Agent 足够的时间来适应策略的变化。平均策略旨在取代未知的其他 Agent 均衡策略,它和当前策略的不同被用作确定算法赢输的标准。在很多博弈中,平均贪婪策略在实际上是近似均衡策略的,这是均衡发挥作用的驱动机制。WoLF 准则使得 PHC 算法在自身博弈中可以收敛至纳什均衡解。由此,该算法在保留理性的基础上增加了收敛的性

质,使其能收敛至其中的一个纳什均衡解。收敛性质将从下文的几个典型案例计算来开展。Agent i 的 Q 学习更新规则如下:

$$Q_i^{t+1}(s,a)=(1-a_t)Q_i^t(s,a)+a_t[r_i^t+\gamma \mathrm{Max}Q_i^t(s',a')] \tag{8}$$

据此,设计 WoLF-PHC 算法如下所示:

步骤一:初始化 $Q^i(s,a_i)=0$,$\pi^i(s,a_i)=\frac{1}{|A_i|}$,$C(s)=0$,选择学习率 α,δ,以及折扣率 γ。

步骤二:对每一次迭代,

1) 在当前状态下基于一个混合探索-利用策略选择行动 a_c,执行行动后观察收益 r_i 以及下一个状态 s';

2) 更新 $Q_i(s,a_c)$:$Q_i(s,a_c)=Q_i(s,a_c)+\alpha[r_i+\gamma \max\limits_{a'_i} Q(s',a_i')-Q(s,a_c)]$,其中 a_i' 是玩家 i 在下一个状态 s' 时的行动,a_c 是玩家 i 在状态 s 时的行动;

3) 更新平均策略 $\overline{\pi_i}$ 的估值:

$C(s)=C(s)+1$,

$$\overline{\pi_i}(s,a_i)=\overline{\pi_i}(s,a_i)+\frac{1}{C(s)}(\pi_i(s,a_i)-\overline{\pi_i}(s,a_i)),\ \forall a_i\in A_i,$$

其中 $C(s)$表示状态 s 被访问过的次数;

4) 更新 $\pi_i(s,a_i)$,$\pi_i(s,a_i)=\pi_i(s,a_i)+\Delta_{sa}$, $\forall a_i\in A_i$,其中

$$\Delta_{sa}=\begin{cases}-\delta_{sa} & \text{if } a_c\neq \arg\max_{a_i\in A_i} Q_i(s,a_i)\\ \sum_{a\neq a} sa_j & \text{otherwise}\end{cases}$$

$$\delta_{sa}=\min(\pi_i(s,a_i),\frac{\delta}{|A_i|-1})$$

$$\delta=\begin{cases}\delta_w & \text{if } \sum_{a\neq A_i}\pi_i(s,a_i)Q_i(s,a_i)>\sum_{a\in A_i}\overline{\pi}_i(s,a_i)Q_i(s,a_i)\\ \delta_l & \text{otherwise}\end{cases};$$

步骤三:更新 $S=S'$,若 S 为最终状态,结束该循环;否则,返回至步骤二。

此算法的状态 S 所包含的剩余座位数变量的设置同 Nash-Q 算法一致。

3　仿真分析

假设两个航班的参数一致。可销售的票价集为{10,15,20,25,30},总座位数均为 50。假设剩余销售期为 10 天,对其进行离散化处理,使得每天为一个时间段。状态包括两个航班的剩余座位数以及当前的时间段。其中,变量 x_i^t,将每 5 个座位数为一个状态,即将剩余舱位数除以 5 取整作为当前的剩余舱位数状态。对旅客仿真环境中的参数进行设置,建立仿真环境。参数设置为:$\varpi=0.4$,$\varphi_s=0.6$,$\varphi_m=0.6$,$\beta_{LS}=0.2$,$\beta_{HS}=0.1$,不同类型旅客设定不同的 λ,υ_i,η。Nash-Q 算法中的参数设置为:$\alpha_i=0.1$,$\gamma=0.9$。WoLF-PHC 算法中的参数设置为:$\alpha_i=0.1$,$\gamma=0.9$,$\delta=0.000\ 1$。

通过案例计算,Nash-Q 算法在 5 000 次左右收敛,WoLF-PHC 算法在 100 000 次左右收敛,表明两种算法在实际问题中都具有较好的收敛性。Nash-Q 算法收敛所需的迭代次数要少于 WoLF-PHC 算法,但 WoLF-PHC 算法的收敛速度明显优于 Nash-Q 算法。这是符合实际情况的,因为 Nash-Q 算法在每个阶段中都会计算纳什均衡解,而 WoLF-PHC 算法只能利用平均策略来近似策略,这使得其需要更多的迭代后才能收敛。而也正因为 Nash-Q 算法需要在每一次阶段博弈时计算纳什均衡解,这会大大增加求解时间,且该算法需要已知自身和竞争对手双方的 Q 值,这对于存储空间的要求也有所增加。因此,WoLF-PHC 算法不论是在求解时间还是耗费的存储空间上都要优于 Nash-Q 算法。

表 1 是航班在各个时间段的定价。索引行为距离离港日期的时间,0 为停止售票的时间点。Strategy 1 为旅客的保留价格稳定不变时的定价策略,Strategy 2 为航班在旅客保留价格随时间而变时的定价策略。

Strategy 3 和 Strategy 4 考虑策略型旅客及保留价格变化的定价策略，且 Strategy 4 的旅客策略程度的参数设置的要高于 Strategy 3，由此可以看出策略程度对价格制定也会产生一定的影响，使得整体制定的价格都稍低一些。从整体上看，由于航空旅客到达的特点，机票的价格曲线处于一个增长的趋势。但由于旅客保留价格的变化及其策略行为使得航空公司需要不断的调整价格，这也就导致价格在增长的过程中会出现一些波动，这些波动能更好地适应旅客行为的变化。将定价策略放到仿真环境中模拟，得到的收益相比于传统方法提升约 1.34%。

表 1 航班定价策略

Tab.1 The pricing strategy of flight tickets 元

时间/天	定价策略			
	Strategy 1	Strategy 2	Strategy 3	Strategy 4
10	10	10	10	10
9	10	15	15	15
8	15	20	15	20
7	15	15	20	15
6	15	20	25	25
5	15	25	25	20
4	20	20	30	25
3	20	25	25	20
2	20	30	20	20
1	25	25	30	25
0	30	30	25	30

4 结论

利用强化学习算法求解平行航班的动态定价问题，发现该算法对多主体的动态定价问题具有较好的适应性，且能在有限步骤内得到收敛，计算时间相比于传统的近似计算方法较短且更加精确。通过 Nash-Q 算法和 WoLF-PHC 算法的求解结果比较，发现 WoLF-PHC 算法在求解时间上都优于 Nash-Q 算法，且 Nash-Q 算法在维护 Q 值表上耗费的空间较大。此外，WoLF-PHC 算法在不同的旅客环境中，定价策略也会发生相应的变化，能较好地适应旅客环境的变化。由于航空旅客到达策略与其他行业有所不同，高消费的旅客往往到出发前期才会购买机票，而低消费的旅客会早早的选择进行购票，使得航空机票在整体上呈现出增长趋势。另外，由于旅客的策略性行为，以及保留价格分布的变化，这也使得航空机票在定价过程中会出现一些波动以适应旅客的随机行为。WoLF-PHC 算法在平行航班的动态定价问题中的表现优于 Nash-Q 算法，且得到的定价策略能有效地增加航空公司收益，对航空公司在越来越剧烈的竞争市场中立于不败之地具有重要作用。

参考文献：

[1] 顾颖菁，周海花. 基于多阶段动态组合拍卖的联盟舱位分配研究[J]. 华东交通大学学报，2018，35(6)：48-54.

[2] 李金林，雷俊丽，冉伦，等. 航空收益管理柔性舱位控制机制的研究现状与展望[J]. 北京理工大学学报，2012，32(4)：331-347.

[3] 高金敏，乐美龙，曲林迟，等. 机票定价与舱位控制两阶段方法[J]. 控制与决策，2019，34(6)：175-181.

[4] GALLEGO G，GARRETT V R. A multiproduct dynamic pricing problem and its applications to network yield management[J]. Operations Research，1997，45(1)：24-41.

[5] ZHANG D,COOPER W L. Pricing substitutable flights in airline revenue management[J]. European Journal of Operational Research,2009,197(3):848-861.
[6] AKCAY Y,NATARAJAN H P,XU S H. Joint dynamic pricing of multiple perishable products under consumer choice[J]. Management Science,2010,56(8):1345-1361.
[7] GALLEGO G,HU M. Dynamic pricing of perishable assets under competition[J]. Social Science Electronic Publishing,2014,60(5):1241-1259.
[8] 曹海娜. 基于 MNL 模型的平行航班舱位控制与动态定价研究[D]. 北京:北京理工大学,2015.
[9] 朱志愚,王宗宝,刘燕,等. 竞争环境下多价格等级的平行航班动态定价模型研究[J]. 科技和产业,2016,16(9):106-112.
[10] HAN W,LIU L,ZHENG H. Dynamic pricing by multiagent reinforcement learning[C]//International Symposium on Electronic Commerce & Security,IEEE,2008.
[11] 王金田,唐昊,程文娟,等. 基于强化学习的异步动态定价算法[J]. 系统工程学报,2011,26(5):664-670.
[12] 陆慧,基于多 Agent 的季节性商品动态定价算法[J]. 计算机应用,2011,31(11):3135-3139.
[13] RANA R,OLIVEIRA F S. Real-time dynamic pricing in a non-stationary environment using model-free reinforcement learning[J]. Omega,2014,47(9):116-126.
[14] RANA R,OLIVEIRA F S. Dynamic pricing policies for interdependent perishable products or services using reinforcement learning[J]. Expert Systems with Applications,2015,42(1):426-436.
[15] KIM B G,ZHANG Y,SCHAAR M V D,et al. Dynamic pricing for smart grid with reinforcement learning[C]//Computer Communications Workshops,IEEE,2014.
[16] LU R,HONG S H,ZHANG X A Dynamic pricing demand response algorithm for smart grid: Reinforcement learning approach[J]. Applied Energy,2018,220:220-230.
[17] MA Q,MENG F,ZENG X J. Optimal dynamic pricing for smart grid having mixed customers with and without smart meters[J]. Journal of Modern Power Systems and Clean Energy,2018,6:1244-1254.
[18] SCHWARTZ H M. Multi-agent machine learning: a reinforcement approach[M]. Wiley Publishing,2014:144-199.

Dynamic Pricing of Parallel Flights Based on the Reinforcement Learning

Fang Yuan,Le Meilong

(College of Civil Aviation, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract:The competition between parallel flights is becoming increasingly fierce. In this study, to improve the airline's revenue, the flights and the passengers were separately modeled in the ticket sale system. The problem of dynamic pricing of flights was modeled as Markov game, and the Logit choice model was used to model for the mixed-type passengers. The multi-agent reinforcement learning was adopted to solve the problem in reality. The results indicated that the number of convergence for WoLF-PHC algorithm was more than that of the Nash-Q, but the WoLF-PHC algorithm had higher convergence frequency with strong adaptability. In addition, the pricing strategy of flight ticket sale process was different from that of other perishable products, which generally reflected an upward trend. The pricing strategy would also be adjusted with the modification of environment parameters of passengers. The pricing policy obtained by WoLF-PHC algorithm has positive effects on improving revenue.

Key words: paralle lflights;mixed-type passengers;dynamic pricing;Markov game;reinforcement learning