· 方法介绍 ·

基于强化学习确立动态治疗方案的方法介绍*

南方医科大学公共卫生学院生物统计学系(510515) 梁雪晴 余世杰 朱思宇 罗熠欣 吴 莹△ 段重阳△

【提 要】 随着医学治疗越来越趋向于精准化、个体化,如何有效利用病人动态更新的信息成为现代医学关注的一大问题,由此学者提出了充分考虑患者异质性和长期治疗效益的动态治疗方案(dynamic treatment regimes)。动态治疗方案是一组顺序决策规则,每条规则对应于疾病发展的一个关键点,根据患者的累计信息决定下一次治疗。强化学习方法,特别是其中的 Q-learning 方法,因其特有的模型特点,被广泛应用于探索动态治疗方案。本文旨在详细介绍动态治疗方案下的强化学习框架和 Q-learning 方法的应用,并分别就连续型结局和生存结局数据用 R 软件操作举例说明。

【关键词】 动态治疗方案 强化学习 Q-learning 生存数据

【中图分类号】 R195.1 【文献标识码】 A **DOI** 10.11783/j.issn.1002-3674.2024.05.036

随着现代化医疗的不断发展,治疗越来越趋向于精准化、个体化,动态治疗方案(dynamic treatment regimes)的概念也随之被提出[1]。动态治疗,意为治疗是一个会随时间发生改变的变量,而动态治疗方案,是指一组连续的决策规则,其中每条规则对应于疾病或疾病发展的一个关键点,在这个关键点上医生必须对病人的下一步治疗行动做出决定[2]。现在许多疾病的治疗是分阶段进行的,如对非小细胞肺癌患者的治疗方案通常为三线治疗[3]。对于这类疾病,医生须在特定的阶段根据病人当前时间点的病史和接受的过往治疗来选择最优治疗方案。而最优治疗方案就是在最后的治疗结束后使某些临床结果的平均反应最大化(或最小化),即最优的一组治疗选择[4]。与所有个体被分配相同水平和治疗类型的经典治疗相比,动态治疗考虑了个体间治疗需求的异质性和个体内跨时间的治疗需求异质性[1],对于个体治疗方案的选择而言更为灵活可靠。

动态治疗方案概念的提出弥补了传统设计忽略个体异质性的缺点,但也带来了方法研究上的困难与挑战。如何将长期积累的信息纳入决策并选出长期效果最优的治疗方案成为了一大难题,而强化学习作为第一个解决从与环境的互动中学习以实现长期目标时出现的计算问题的领域[5],被尝试应用于探索动态治疗方案。强化学习是一种解决多阶段决策问题的方法,它包括记录行动序列,统计并估计这些行动和它们造成的结果之间的关系,然后根据统计估计结果选择一个最接近理想结局的决策(即一套决策规则)[6]。在临床应用方面,强化学习已被应用于治疗行为障碍,这种疾病的病人通常有多次机会尝试不同的治疗方法[7]。Murphy 等人[8]建议将 Q-learning 用于构建慢性精神疾病的决策规则,因为这些慢性疾病往往需要医生通过连续的决策以达到最佳的临床效果。Moodie 等人先后提出将传统的 Q-learning 方法拓展到考虑纳入测量的混杂协变量[9],以及使用广义加性模型而不是线性模型来估计 Q 函数[10]。而 Zhao 等[3,11]首次提出将 Q-learning 应用于危及生命的疾病如癌症。Goldberg 和 Kosorok[4]在 Q-learning 的基础上对生存数据中删失的处理做出了一些拓展。Simoneau 等[12]则在 Q-learning 的思路基础上提出了一种新的、理论上稳健的方法来估计右删失的生存数据的最优动态治疗方案。这些方法的提出为动态治疗方案的研究提供了良好的理论基础。

在这篇文章中,我们首先介绍通用的强化学习框架以及其在医学背景中动态治疗决策领域下的应用。其次介绍强化学习中的重要算法 Q-learning 方法,然后详细介绍处理生存结局的动态治疗决策方法。最后使用 R 软件包介绍连续型结局和生存结局数据的动态治疗方案选择的实例分析。

动态治疗决策下的强化学习框架

在医疗背景下,强化学习的设置可以表示如下。对于每个病人,各阶段对应于病人治疗过程中的临床决策点。在这些决策点,病人接受不一样的治疗(即采取行动 A_t),并且记录下病人的各项身体指标(即所处状态 S_t)。在接受治疗后,病人会发生病情的变化(即得到奖励 R_t)[4]。

为了更详细地介绍以上过程,我们以有两个关键决策点的急性白血病患者的治疗方案选择问题为例说明。其中决策 1 为选择一种诱导化疗,旨在诱导预先指定的反应,如完全或部分缓解。假设有一组两个诱导方案,$A_1=\{B_1,B_2\}$。诱导治疗完成后,决策 2 包括为有响应的患者选择维持治疗(M_1,M_2),或为没有响

* 基金项目:国家自然科学基金面上基金项目(82273727);国家自然科学基金青年科学基金项目(81803327)

△通信作者:吴莹,E-mail:wuying19890321@ gmail.com;段重阳 E-mail:donyduang@ 126.com

应的患者选择挽救治疗(S_1,S_2)。由于有两个维持治疗方案和两个挽救治疗方案,决策 2 的选项集为 $A_2=\{M_1,M_2,S_1,S_2\}$。S_1 为决策点一时可获得的病人信息,该例使用年龄(岁)和白细胞计数(WBC×10^3 μL^{-1})表示。此时,若历史信息 h_1 包括 age 和 WBC,决策 1 的简单规则可表示为 $d_1(h_1)=I(age<50\ and\ WBC<10)$ 或 $d_1(h_1)=I\{age+7.8log(WBC)-60>0\}$,其中 $I(\cdot)$ 为指示函数。在决策 2 中,可行的规则 $d_2(h_2)$ 必须取决于反应状态,并且只返回反应者(无反应者)的维持(挽救)选项。

下面使用通用符号介绍决策过程,假设一个有 T 个决策点的多阶段决策问题[3]。S_t 表示病人在阶段 $t\in\{0,\cdots,T\}$ 时的状态,并定义直到并包括阶段 t 的所有状态的向量 $S_t=\{S_0,S_1,\cdots,S_t\}$。相似地,$A_t$ 表示在阶段 t 采取的行动,$A_t=\{A_0,A_1,\cdots,A_t\}$ 是到阶段 t 为止的所有行动的向量。变量 S_t 和 A_t 都可以是连续或离散的。我们还可以将状态定义为可能包括过去的行动。使用小写字母 s 和 a 来表示 S 和 A 的观测值,同时,定义 $s_t=\{s_0,s_1,\cdots,s_t\}$ 和 $a_t=\{a_0,a_1,\cdots,a_t\}$。我们用 $R_t=r(S_t,A_t,S_{t+1})$ 来表示病人在阶段 t 获得的随机奖励,其中 r 是直到阶段 $t+1$ 的所有状态和直到阶段 t 的所有过去动作的(未知的)时间相关的确定性函数。强化学习就是学习如何将情况从状态空间 S 映射到行动空间 A,并根据我们的目标来选择最大化或最小化预期折扣回报(expected discounted return):

$$\tilde{r}_t=r_t+\gamma r_{t+1}+\gamma^2 r_{t+2}+\cdots+\gamma^T r_{t+T}=\sum_{k=0}^{T-t}\gamma^k r_{t+k} \tag{1}$$

这里 γ 表示折扣因子($0\leqslant\gamma\leqslant1$),其用于平衡病人的即时奖励(immediate rewards)和未来奖励(future rewards)。当 γ 接近 1 时,我们将给予未来回报更高的权重。在极端情况下,当 $\gamma=1$ 时,我们给未来的回报和即时回报赋予同等的权重。

强化学习系统的另一个关键因素是探索“策略”,p,它将(s_t,a_{t-1})映射到概率 $p_t(a|s_t,a_{t-1})$(在历史$\{s_t,a_{t-1}\}$下采取行动 a 的概率)。我们令 $\pi_t(s_t,a_{t-1})=a_t$,即策略 π_t,作为决策规则$\{\pi_1,\cdots\cdots,\pi_T\}$的序列,是一个行动。使用分布$P_\pi$ 表示训练数据的分布,当策略 π 被用来生成行动时,我们可以用 E_π 表示相对于分布 P_π 的期望。使用 Π 表示所有策略的集合。那么,期望 E_π 的范围是 $\pi\in\Pi$。简单起见,我们在本文中主要聚焦于发现哪种治疗方法能给特定病人带来最大的回报这一目标,也就是寻求使时间轨迹上的奖励之和的期望值最大化的策略。为了实现这一目标,我们建立了“价值函数”作为状态和行动的函数。基于历史条件,价值函数代表了病人在未来可以期望积累的奖励总量,即:

$$V_t(s_t,a_{t-1})=E_\pi\Big[\sum_{k=0}^{T-t}\gamma^k R_{t+k}\mid S_t=s_t,A_{t-1}=a_{t-1}\Big] \tag{2}$$

据此,最优价值函数可以简单定义为:

$$V_t^*=\max_{\pi\in\Pi}V_t(s_t,a_{t-1})=\max_{\pi\in\Pi}E_\pi\Big[\sum_{k=0}^{T-t}\gamma^k R_{t+k}\mid S_t=s_t,A_{t-1}=a_{t-1}\Big] \tag{3}$$

有效地估计最佳价值函数是几乎所有强化学习算法最重要的组成部分。由于整个强化学习中使用的价值函数的一个基本属性是它们满足特定的递归关系,如贝尔曼方程(Bellman equation)[13],很明显,最优策略π^*必须满足:

$$\pi_t^*(s_t,a_{t-1})\in\arg\max_{a_t}E\big[R_t+\gamma V_{t+1}^*(S_{t+1},A_t)\mid S_t=s_t,A_t=a_t\big] \tag{4}$$

强化学习的主要目标是找到一种能够带来高的累积期望回报(high expected cumulative reward)的策略。简单地说,我们可以使用观察到的轨迹来学习转移分布函数(transition distribution functions)和奖励函数(reward function),然后递归地求解 Bellman 方程[13]。然而,这种方法的计算量大,计算机运行效率偏低[14]。下面我们将着重介绍的 Q-learning 算法,该算法的计算量小,目前是最有效且最常见的不依赖模型的强化学习算法[4-5]。

Q-learning

Q-learning[15]是一种解决强化学习问题的算法。Q-learning 不直接估计价值函数,而是估计一个 Q 函数。它不需要先验知识,不需要假设参数并且容易实现[3]。在我们没有任何关于转移函数或随机变量的概率分布信息的情况下,使用这样不依赖模型的方法可以从未知系统中找到最优策略。

我们定义最优时变 Q 函数(optimal time-dependent Q-function)为:

$$Q_t^*(s_t,a_t)=E\big[R_t+\gamma V_{t+1}^*(S_{t+1},A_t)\mid S_t=s_t,A_t=a_t\big]$$

这里的 $V_t^*(s_t,a_{t-1})=\max_{a_t}Q_t^*(s_t,a_t)$,因此,

$$Q_t^*(s_t,a_t)=E\big[R_t+\gamma\max_{a_{t+1}}Q_{t+1}^*(S_{t+1},A_t,a_{t+1})\mid S_t=s_t,A_t=a_t\big] \tag{5}$$

为了估计最优策略,我们首先通过时间 $t=T,T-1,\cdots,1,0$ 来反向估计 Q 函数,并得到一个估计值序列$\{\widehat{Q}_T,\cdots,\widehat{Q}_0\}$。所估计的策略为:

$$\widehat{\pi}_t(s_t,a_{t-1})=\underset{a_t}{argmax}\,\widehat{Q}_t(s_t,a_{t-1},a_t) \tag{6}$$

单步(one-step)Q-learning 具有简单的递归形式:

$$Q_t(s_t,a_t)=E\big[R_t+\gamma\max_{a_{t+1}}Q_{t+1}(S_{t+1},A_t,a_{t+1})\mid S_t=s_t,A_t=a_t\big] \tag{7}$$

一组轨迹(one set of finite horizon trajectories)(也称为训练数据集)被定义为:

$$\{S_0,A_0,R_0,S_1,A_1,R_1,\cdots,A_T,R_T,S_{T+1}\}$$

我们用$\widehat{Q}_t$来表示基于该训练数据的最佳 Q 函数的估计，其中 $t=0,1,\cdots,T$。根据(5)中 Q-learning 的递归形式，我们必须通过时间 $t=T,T-1,\cdots,1,0$ 逆向估计 Q_t，也就是说，使用从最后一个时间点$\widehat{Q}_T$的估计值递归到最开始的$\widehat{Q}_0$。为方便起见，我们设定$\widehat{Q}_{T+1}$等于 0。为了估计每个 Q_t，我们把 $Q_t(s_t,a_t;\theta_t)$定义为一组参数为 θ 的函数，并且允许该函数在不同的时间点 t 有不同的参数集。例如，$Q_t(s_t,a_t;\theta_t)$的形式可以是

$$Q_t(s_t,a_t;\theta_t)=\sum_{j=1}^{k}\theta_{tj}\varphi_{tj}(s_t,a_t)$$

其中 $\theta_t=(\theta_{t1},\cdots,\theta_{tk})$和$\{\varphi_{t1},\cdots,\varphi_{tk}\}$是选定的基础函数。一旦逆向估计过程完成，我们将使用估计出的序列$\{\widehat{Q}_0,\widehat{Q}_1,\cdots,\widehat{Q}_T\}$来估计最优策略。

$$\widehat{\pi}_t=\underset{a_t}{argmax}\widehat{Q}_t(s_t,a_t;\theta_t)$$

其中 $t=0,1,\cdots,T$，此后我们使用这些最优策略来测试或预测新的数据集。

接下来，我们的主要目的是估计 Q 函数以找到相应的最优策略。然而，由于真实 Q 函数结构的复杂性，包括方程(5)中的非光滑最大化估计，状态变量 S 和动作变量 A 的高维度属性，或是存在连续型的动作变量 A 等，都会对估计 Q 函数带来巨大挑战。为了获得感兴趣的估计值，近年来许多作者考虑了不同的方法。

Murphy 以及 Tsitsiklis 和 Van Roy 等[16-18]表示，Q-learning 估计可以被看作是近似的最小二乘值迭代。第 t 个 Q 函数的参数估计值$\widehat{\theta}_t$满足

$$\widehat{\theta}_t\in\arg\min_{\theta}\mathbb{E}_n$$
$$[R_t+\max_{a_t+1}\widehat{Q}_{t+1}(S_{t+1},a_{t+1};\widehat{\theta}_{t+1})-Q_t(S_t,A_t;\theta)]^2$$

其中$\mathbb{E}_n$是经验期望值。Murphy 等[19]提供了另一种简单而标准的估计形式，他们认为 Q-learning 是回归模型的推广。也有许多学者将线性形式的 Q-function 应用于探索连续型结局的最优动态治疗方案[8,19-20]。Moodie[9]提出使用直接调整和各种倾向评分方法(包括回归和逆概率加权)，将传统的 Q-learning 方法拓展到考虑纳入测量的混杂协变量的情形。此外，为了更好地反映 Q 函数的复杂性和结构不明确性，Zhao[3]提出使用两种机器学习的方法，即支持向量回归方法(support vector regression)和极度随机树(extremely randomized trees)，来估计 Q 函数。Moodie 等[10]提出使用广义加性模型来估计 Q 函数，并将此方法应用于估计连续型结局和二分类结局的最优治疗策略。在软件应用方面，基于回归的 Q-learning 方法目前已通过 R 包 qLearn[21]得到实现。Laber 和 Linn[22]提出了 Q-learning 的一种替代方法 Interactive Q-learning 以及对应的 R 包 iqLearn[23]，该方法只需要对数据的平滑、单调的变换进行建模。同时，一些更复杂的、基于分类的方法[24-27]在 R 包 DynTxRegime[28]中得以实现。Wallace 等[29]编写了 R 软件包 DTRreg，包中实现了 G 估计(G-estimation)，动态加权普通最小二乘法(dynamic weighted ordinary least squares)以及简单设定下的 Q-learning 方法(只限于更简单的、两段式的二元处理设置，不能取代现有的 qLearn 软件包)。由于方法众多，此处不详细讲解，仅在后续实例说明中使用 DTRreg 中的 G-estimation 方法来选择连续型结局的最优动态治疗方案。

处理生存数据的 Q-learning

有了关于 Q-learning 的框架知识，我们接下来详细介绍处理生存数据中有删失数据的 Q-learning 方法。我们在估计最优动态治疗制度时纳入删失轨迹的数据会存在一定的问题：首先，即使阶段数是固定的，删失轨迹的已知阶段数可能少于多阶段问题的阶段数。此外，在发生删失的阶段，个体的奖励是不知道的。为了解决以上问题，Zhao 等人[3]考虑了一种基于支持向量回归的有固定阶段数的删失数据的 Q-learning 算法，并对该算法进行了模拟研究以说明其性能。然而，该算法的理论特性没有得到评估。Goldberg 和 Kosorok[4]也在 Q-learning 思路的基础上针对删失数据进行了一定的扩展，其使用逆概率删失权重(inverse probability censoring weighted)来考虑删失问题。它的建模方法简单而直观，然而，它缺乏对模型错误指定的稳健性，而且它假设删失与观察到的轨迹无关，在软件使用上目前其只在 MATLAB 中实现。Simoneau 等人[12]提出了一种新的、理论上稳健的方法来估计右删失的生存数据的最优动态治疗方案。该方法将动态加权普通最小二乘法[29]扩展到了生存数据，并借鉴了 Huang 等人[30]建立的单一稳健框架，是一种非常容易实现的统计方法，并且当结果是连续的和无删失的时候，它结合了 G 估计的双重稳健性和 Q-learning 的简单性。其针对生存数据的扩展，主要包括一系列灵活的干预阶段数，从而允许个体在随访结束前经历事件或被删失，并且允许删失依赖于时变的个体轨迹。该方法被命名为动态加权生存模型(dynamic weighted survival modeling，DWSurv)，其在理论上是双重稳健的，在软件实现上，可以直接调用 R 中 DTRreg 包的 DWSurv 函数[30]。考虑其性能和操作的优越性，接下来详细介绍 DWSurv 方法。

1.动态加权生存模型

(1)符号定义与假设

我们使用大写字母表示随机变量，小写字母表示随机变量的观察值。假设一个 J 阶段的临床干预数据，我们使用下标 $i=1,\cdots,n$ 来表示个体，下标 $j=1,\cdots,J$ 来表示阶段。定义随机变量 η_j，若个体进入阶段

j 则 $\eta_j=1$，否则为 0，所有个体的 $\eta_1=1$。使用 $A_j\in(0,1)$ 来表示在阶段 j 的开始是否接受治疗，T_j 为阶段 j 内的生存时间，当 $\eta_j=0$ 时 $T_j=0$。定义结局为总的生存时间，$T=\sum_{j=1}^{J}\eta_j T_j$。同时，我们定义反事实生存时间 $T^a=\sum_{j=1}^{J}\eta_j T_j^{a_j}$，其中 $T_j^{a_j}$ 表示其在阶段 j 接受治疗 a_j 的潜在生存时间。使用 C 来表示删失时间。令观察到的个体生存时间 $Y=min(T,C)$，删失的指示变量为 $\Delta=I(T\leqslant C)$。向量 X_j 为第 j 个治疗前测量的协变量，H_j 为第 j 个治疗决策前个体的历史信息，包括以前的治疗分配、协变量和生存时间的函数。观察数据由个体的轨迹 $(\eta_{i1},X_{i1},A_{i1},T_{i1},\cdots,\eta_{iJ},X_{iJ},A_{iJ},T_{iJ},\Delta_i)$ 组成，当 $\eta_{ij}=0$ 时 A_{ij} 和 X_{ij} 为缺失值。一个动态治疗方案由一组决策规则 $\{d(h_1),\cdots,d(h_J)\in D\}$ 组成，其中 D 表示所有可能的治疗策略的类别。在每个阶段 j，决策规则是一个函数 $d(h_j):H_j\rightarrow(0,1)$，根据观察到的历史 h_j 获得一个治疗方案。一个最佳的动态治疗方案是使总体预期生存时间 $E(T^a)$ 最大化得到的决策规则的集合 $\{d^{opt}(h_1),\cdots,d^{opt}(h_J)\}$。

为了确定一个最佳的动态治疗方案，我们依靠一致性公理将反事实与观察数据联系起来。做出以下假设：

①稳定的单位治疗值（stable unit treatment value）[31]：要求一个个体的结局不受其他个体治疗分配的影响。

②序列可忽略性（sequential ignorability）[32]：这是扩展到纵向设置（longitudinal settings）的没有未测量的混杂因素（unmeasured confounder）假设，这进一步要求在给定阶段的治疗分配不能依赖于未来的协变量。可表示为 $\{\sum_{k\geqslant l}^{J}T_k^{a_k}:l=j,\cdots,J\}\perp\!\!\!\perp A_j\mid H_j,\eta_1\cdots,\eta_j$。

③随机删失（censoring at random）[33]：它假定，在每个阶段的开始，考虑到累积的信息，往后删失的概率与未来的结局无关。可表示为 $\{\sum_{k\geqslant l}^{J}T_k^{a_k}:l=j,\cdots,J\}\perp\!\!\!\perp\Delta\mid H_j,\eta_1\cdots,\eta_j$。

（2）方法介绍

为简单起见，我们定义了一个最多有两个干预阶段的最优动态治疗方案，并用简写 $a_1{}^{opt}$ 和 $a_2{}^{opt}$ 分别表示阶段 1 和阶段 2 的最优决策规则 $d^{opt}(h_1)$ 和 $d^{opt}(h_2)$。将下面的推导和结果扩展到两个或更多阶段是很容易的。

首先，通过使第二阶段生存时间的对数最大化来估计第二阶段最优治疗 $a_2{}^{opt}$。其次，通过在接受第二阶段最优治疗的情况下使总体生存时间对数最大化来估计第一阶段最优治疗方法 $a_1{}^{opt}$。这种反事实的结果可以使第一阶段的治疗得到公平的比较，因为通过反事实结局我们可以在假设所有进入第二阶段的受试者都接受了第二阶段最优治疗的基础上，使总生存时间对数最大化以估计第一阶段的最优治疗。具体算法介绍如下：

①阶段二最优化

1）构建第二阶段治疗概率的模型 $P[A_{i2}=1\mid h_{i2},\eta_{i2}=1;\alpha_2]$（the treatment model）以及第二阶段发生事件的概率的模型 $P[\Delta_i=1\mid h_{i2},\eta_{i2}=1;\lambda_2]$（the censoring model）。使用进入第二阶段的受试者拟合这两个模型（$\eta_{i2}=1$），例如，使用逻辑回归。

2）对第二阶段的生存时间构建一个半参数 AFT 模型：

$$\mathbb{E}[\log(T_2{}^{a_1,a_2})\mid h_2,a_2;\beta_2,\psi_2]=f_2(h_{2\beta};\beta_2)+a_2\,g_2(h_{2\psi};\psi_2)+\epsilon_2$$

其中，误差 ϵ_2 在不同个体间是独立同分布的。$\log(T_2{}^{a_1,a_2})$ 的模型可分为两部分：无治疗部分（treatment-free component）$f_2(h_{2\beta};\beta_2)$，这部分受阶段二一部分历史 h_2（不包括治疗 a_2）的影响，代表了在没有治疗的情况下自然疾病发展对阶段二的生存时间的影响。以及阶段二的治疗效果部分（treatment effect component）$a_2\,g_2(h_{2\psi};\psi_2)$，这部分取决于（一个可能不同的子集）第二阶段的历史，具体包括治疗 a_2 的主要影响及其与裁剪协变量（tailoring covariates）的相互作用。其中 f_2 和 g_2 可以使用任何函数形式，参数化中最典型的选择为线性回归：

$$\mathbb{E}[\log(T_2{}^{a_1,a_2})\mid h_2,a_2;\beta_2,\psi_2]=\beta_2^Th_{2\beta}+a_2\psi_2^Th_{2\psi}\quad(8)$$

3）通过求解加权估计方程来估计参数 (β_2,ψ_2)：

$$U_2(\psi_2,\beta_2)=\sum_{i=1}^{n}\delta_i\eta_{i2}\widehat{\omega}_{i2}\begin{pmatrix}h_{i2\beta}\\a_{i2}h_{i2\psi}\end{pmatrix}(\log(T_{i2})-\mathbb{E}[\log(T_2{}^{a_1,a_2})\mid h_2,a_2;\beta_2,\psi_2])=0\quad(9)$$

使用满足平衡特性（balancing property）的权重，如：

$$\widehat{\omega}_{i2}=\frac{\mid a_{i2}-\mathbb{E}[A_{i2}\mid h_{i2},\eta_{i2}=1;\widehat{\alpha}_2]\mid}{\mathbb{P}(\Delta_i=\delta_i\mid h_{i2},\eta_{i2}=1;\widehat{\lambda}_2)}$$

4）估计第二阶段最优治疗：

最佳的第二阶段治疗规则是使 blip 模型取值最大化，阶段二的 blip 模型为：

$$\gamma_2(a_2,h_2;\psi_2)=\mathbb{E}[\log(T_2{}^{a_1,a_2})-\log(T_2{}^{a_1,0})\mid\eta_2=1,H_2=h_2;\psi_2],$$

将式（8）代入可得：

$$\gamma_2(a_2,h_2;\psi_2)=\mathbb{E}[\log(T_2{}^{a_1,a_2})-\log(T_2{}^{a_1,0})\mid\eta_2=1,H_2=h_2;\widehat{\psi}_2]$$

$$=\widehat{\beta}_2^Th_{2\beta}+a_2\widehat{\psi}_2^Th_{2\psi}-(\widehat{\beta}_2^Th_{2\beta}+0\times\widehat{\psi}_2^Th_{2\psi})$$

$$=a_2\widehat{\psi}_2^Th_{2\psi}$$

故，最优治疗为 $\widehat{a_2}^{opt}=\mathbb{I}(\widehat{\psi}_2^Th_{2\psi}>0)$。

5）构建在第二阶段最优治疗下的反事实生存时间 $\widetilde{T}$：

$$\widetilde{T}(\widehat{\psi}_2)=T_1+\eta_2(T_2\times\exp\{\widehat{\psi}_2^T h_{2\psi}[\widehat{a}_2^{opt}-a_2]\})$$
$$=\begin{cases}T_1, if \quad \eta_2=0\\ T_1+T_2, if \quad a_2=\widehat{a}_2^{opt}\\ T_1+T_2\exp(|\widehat{\psi}_2^T h_{2\psi}|), if \quad a_2\neq\widehat{a}_2^{opt}\end{cases}$$

②阶段一最优化

1)构建第一阶段治疗概率的模型 $P[A_{i1}=1|h_{i1};\alpha_1]$(the treatment model)以及第一阶段发生事件的概率的模型 $P[\Delta_i=1|h_{i1};\lambda_1]$(the censoring model)。

2)针对伪生存时间(pseudo-survival time)提出一个半参数 AFT 模型:

$$\mathbb{E}[\log(\widetilde{T}(\widehat{\psi}_2))|h_1,a_1;\beta_1,\psi_1]=\beta_1^T h_{1\beta}+a_1\psi_1^T h_{1\psi}$$

3)通过求解加权估计方程来估计参数(β_1,ψ_1):

$$U_2(\psi_1,\beta_1;\widehat{\psi}_2)=\sum_{i=1}^{n}\delta_i\,\widehat{\omega}_{i1}\begin{pmatrix}h_{i1\beta}\\ a_{i1}h_{i1\psi}\end{pmatrix}(\log(\widetilde{T}_i(\widehat{\psi}_2))-\mathbb{E}$$
$$[\log(\widetilde{T}(\widehat{\psi}_2))|h_1,a_1;\beta_1,\psi_1])=0$$

4)估计第一阶段最优治疗:

$$\widehat{a}_1^{opt}=\mathbb{I}(\widehat{\psi}^T{}_1 h_{1\psi}>0)。$$

③估计的最佳动态治疗方案是:在第一阶段开始时推荐 $\widehat{a}_1^{opt}$,如果个体进入第二阶段,推荐 $\widehat{a}_2^{opt}$。

实例展示

1.连续型结局数据实例

本节将使用 iqLearn 软件包[23]中的一个名为 bmiData 的模拟数据集进行说明。数据根据一个两阶段的降低身体质量指数(body mass index, BMI)的 SMART 设计模拟而得,其中每个阶段都有两个治疗:代餐(meal replacement, MR)或传统饮食(conventional diet, CD)。bmiData 中的患者特征、治疗方法和结果是基于一项研究代餐对肥胖青少年减肥和降低 BMI 效果的临床试验[34]中收集的一小部分患者协变量。本例中,基线协变量包括性别、种族、父母 BMI 和基线 BMI,时变协变量为四月和十二月的 BMI 值,结果使用第十二个月的 BMI 与基线的 BMI 相比的负百分比变化来表示(较高的数值表明 BMI 下降较多,这是一个理想的临床结果)。具体变量介绍如表 1。

表 1 数据集 bmiData 的变量描述

变量	Support	描述
gender	{0,1}	患者性别,编码女性为 0,男性为 1
race	{0,1}	患者种族,编码非裔美国人为 0,其他为 1
parent_BMI	ℝ	基线时测量的患者父母 BMI
baseline_BMI	ℝ	基线时测量的患者 BMI
A1	{0,1}	第一阶段的治疗,A1=1 为 MR,A1=0 为 CD
month4_BMI	ℝ	第四个月时测量的患者 BMI
A2	{0,1}	第二阶段的治疗,A2=1 为 MR,A2=0 为 CD
month12_BMI	ℝ	第十二个月时测量的患者 BMI

使用上文中介绍的 DTRreg 包中的 DTRreg 方法进行两阶段的治疗策略选择。

R 代码如下:

```
Library(DTRreg)
#治疗
bmiData $ A1<- ifelse(bmiData $ A1 == "MR", 1, 0)
bmiData $ A2<- ifelse(bmiData $ A2 == "MR", 1, 0)
#结局变量(线性结局)
Y<- -100 * (bmiData $ month12_BMIbmiData $ baseline_BMI)
/bmiData $ baseline_BMI
#构建模型
blip.mod<-list(~gender+race+parent_BMI+baseline_BMI,~gender+ parent_BMI + month4_BMI)#blip 模型
treat.mod<- list(A1~ gender + race + parent_BMI + baseline_BMI, A2~ gender + parent_BMI + month4_BMI)#treatment 模型
tf.mod<-list(~gender+ race + parent_BMI + baseline_BMI,~gender + parent_BMI + month4_BMI)#treatment-free 模型
mod<-DTRreg(y, blip.mod, treat.mod, tf.mod, method="dwols", var.estim="bootstrap",data=bmiData)
```

使用 DTRreg 方法拟合模型,结果在阶段一推荐 7.3820 + 2.5141 gender-0.6461 race-1.1408 parent_BMI + 0.7354 baseline_BMI>0 的个体食用代餐,否则接受常规饮食。在阶段二推荐-12.9026-0.9024 gender + 0.3192 parent_BMI + 0.0975 month4_BMI>0 的个体食用代餐,否则接受常规饮食。假设个体接受最优治疗方案的平均结局估计值为 10.04165,相比原始数据的平均结局指标 6.455876 更高。表 2 为该两阶段决策规则所涉及的参数估计及 95%置信区间。

表 2 两阶段最优治疗规则的系数和 95%置信区间

阶段	决策规则中的变量	系数	95% *CI*
阶段 1	Intercept	7.3820	(-7.6051,22.3690)
	gender	2.5141	(-0.7993,5.8275)
	race	-0.6461	(-3.7133,2.4211)
	parent_BMI	-1.1408	(-1.4471,-0.8345)
	baseline_BMI	0.7354	(0.2831,1.1877)
阶段 2	Intercept	-12.9026	(-24.8318,-0.9735)
	gender	-0.9024	(-4.0203,2.2155)
	parent_BMI	0.3192	(-0.0387,0.6772)
	month4_BMI	0.0975	(-0.3904,0.5854)

2.生存结局数据实例

本例参考某关于鼻咽癌动态治疗方案选择的观察性研究,模拟产生数据,并使用上文中介绍的 DWSurv 方法进行单阶段的治疗策略选择。本实例主要研究对于进行诱导化疗后的鼻咽癌患者,如何选择后续治疗方案(放射治疗还是同期放化疗治疗方案)的动态治疗决策模型。我们模拟生成了四个协变量:身体质量指数,血红蛋白(hemoglobin, HGB),乳酸脱氢酶含量(lactate dehydrogenase, LDH)和 C-反应蛋白(C-reactive protein, CRP)。对于每个个体,A_i 表示其接受的治疗方案,其中 $A_i=0$ 表示接受放射治疗,$A_i=1$ 表示接受同期放化疗治疗。Δ_i 表示是否删失($\Delta_i=1$ 为发生事件,0 为删失)。T_i 和 C_i 分别表示生存时间和删失时间。定义真正的最优治疗决策规则为 $I(0.72+0.016BMI+0.0095CRP-0.012HGB+0.002LDH>0)$。数据生成规则如下:

模拟数据样本量:3500

BMI ~ Normal(22.79,3.30)

HGB ~ Normal(121.32,15.10)

LDH ~ Normal(197.31, 52.91)

log(CRP+1) ~ Normal(1.21, 0.87)

A ~ Bernoulli(expit(4.64-0.025BMI-0.014CRP-0.020HGB+0.0001LDH))

Δ ~ Bernoulli(0.16)

ε ~ Normal(0,0.09)

log(T) = 6.74-0.027BMI-0.006CRP+0.007HGB-0.0014LDH + A(0.72 + 0.016BMI + 0.0095CRP - 0.012HGB+0.002LDH)+ε

对于 Δ=0 的个体,C ~ Exp(1/1100)

R 代码如下:

```
Library(DTRreg)
mod<- DWSurv(time=list(~Y),#生存时间
blip.mod=list(~BMI+CRP+HGB+LDH),#blip 模型
treat.mod=list(A ~ BMI+CRP+HGB+LDH), #treatment 模型
tf.mod=list(~BMI+CRP+HGB+LDH),#treatment-free 模型
cens.mod=list(delta ~ 1),#censoring 模型
data=exampledata)
```

使用 DWSurv 方法拟合模型,结果推荐 0.7760 + 0.0103 BMI + 0.0151 CRP - 0.0114 HGB + 0.0019 LDH>0 的个体接受同期放化疗治疗,否则接受放射治疗。结果显示使用 DWSurv 方法估计的最优治疗方案的与真实的最优治疗方案一致率为 97%。假设个体接受最优治疗方案的平均生存时间估计值为 1134.662,相比原始数据的平均生存时间 1056.25 更高。表 3 为该单阶段决策规则所涉及的参数估计及 95%置信区间。

表 3　单阶段最优治疗规则的系数和 95%置信区间

决策规则中的变量	系数	95% *CI*
Intercept	0.7760	0.5189,1.0331
BMI	0.0103	0.0035,0.0172
CRP	0.0151	0.0011,0.0290
HGB	-0.0114	-0.0128,-0.0099
LDH	0.0019	0.0014,0.0023

小　结

动态治疗决策,越来越多的被应用于长期患者护理相关的临床决策,如行为障碍,慢性精神疾病,癌症等[35]。通过对动态治疗决策方法的研究,医生在进行相关治疗方案选择时,可充分考虑不同患者及同一患者在不同时间存在的异质性,并根据与治疗相关的时变协变量,更为科学、客观地制定治疗方案,以提升治疗效果,延长生存时间。

本文围绕动态治疗决策概念,介绍了其主要运用的强化学习思想和强化学习中的重点算法 Q-learning 方法。最后,详细介绍了用于估计右删失生存数据最优动态治疗方案的动态加权生存模型。同时,文中介绍的主要方法均通过 R 软件举例介绍了其软件操作。

总之,随着个性化医疗概念的提出,在临床实践中对医生进行治疗决策的要求不断升高。本文希望通过介绍相关概念,帮助理解动态治疗决策含义及其估计的主要思想。并通过实例介绍,进一步帮助了解动态治疗决策的 R 软件操作。

参　考　文　献

[1] Murphy SA. Optimal dynamic treatment regimes[J]. Journal of the Royal Statistical Society: Series B(Statistical Methodology), 2003, 65(2):331-355.

[2] Davidian M, Tsiatis AA, Laber EB. Optimal Dynamic Treatment Regimes[M]. Wiley StatsRef: Statistics Reference Online. 2016:1-7.

[3] Zhao Y, Kosorok MR, Zeng D. Reinforcement learning design for cancer clinical trials[J]. Statistics in Medicine, 2009, 28(26):3294-3315.

[4] Goldberg Y, Kosorok MR. Q-learning with censored data[J]. Annals of statistics, 2012, 40(1):529-560.

[5] 高阳, 陈世福, 陆鑫. 强化学习研究综述[J]. 自动化学报, 2004(1):86-100.

[6] Sutton RS, Barto AG. Reinforcement learning: An introduction[J]. Robotica, 1999, 17(2):229-235.

[7] Pineau J, Bellemare MG, Rush AJ, et al. Constructing evidence-based treatment strategies using methods from computer science[J]. Drug and Alcohol Dependence, 2007, 88:S52-S60.

(下转封三)

(上接第 800 页)

[8] Murphy SA, Oslin DW, Rush AJ, et al. Methodological Challenges in Constructing Effective Treatment Sequences for Chronic Psychiatric Disorders[J]. Neuropsychopharmacology,2007,32(2):257-262.

[9] Moodie EEM, Chakraborty B, Kramer MS. Q-learning for estimating optimal dynamic treatment rules from observational data[J]. Canadian Journal of Statistics,2012,40(4):629-645.

[10] Moodie EEM, Dean N, Sun YR. Q-Learning: Flexible Learning About Useful Utilities[J]. Statistics in Biosciences,2014,6(2):223-243.

[11] Zhao Y, Zeng D, Socinski MA, et al. Reinforcement Learning Strategies for Clinical Trials in Nonsmall Cell Lung Cancer[J]. Biometrics,2011,67(4):1422-1433.

[12] Simoneau G, Moodie EEM, Nijjar JS, et al. Estimating Optimal Dynamic Treatment Regimes With Survival Outcomes[J]. Journal of the American Statistical Association,2020,115(531):1531-1539.

[13] Bellman R. Dynamic Programming[J]. Science,1966,153(3731):34-37.

[14] 马骋乾, 谢伟, 孙伟杰. 强化学习研究综述[J]. 指挥控制与仿真, 2018,40(6):68-72.

[15] Watkins C, Dayan P. Technical note : Q-learning[J]. Machine Learning,1992,8(3-4):279-292.

[16] Murphy SA. A generalization error for Q-learning[J]. Journal of Machine Learning Research,2005,6:1073-1097.

[17] Blatt D, Murphy S, Zhu J. A-learning for approximate planning[J]. Report, University of Michigan, Ann Arbor, MI,2004.

[18] Tsitsiklis JN, van Roy B. Feature-based methods for large scale dynamic programming[J]. Machine Learning,1996,22(1):59-94.

[19] Chakraborty B, Murphy S, Strecher V. Inference for non-regular parameters in optimal dynamic treatment regimes[J]. Statistical Methods in Medical Research,2010,19(3):317-343.

[20] Nahum-Shani I, Qian M, Almirall D, et al. Q-learning: a data analysis method for constructing adaptive interventions[J]. Psychological Methods,2012,17(4):478-494.

[21] Xin J, Chakraborty B, Laber EB. qLearn: Estimation and inference for Q-learning. R package version 1.0[EB/OL].https://CRAN.R-project.org/package=qLearn.

[22] Laber EB, Linn KA, Stefanski LA. Interactive model building for Q-learning[J] Biometrika,2014,101(4):831-847.

[23] Linn KA, Laber EB, Stefanski LA. iqLearn: Interactive Q-Learning in R[J]. Journal of Statistical Software,2015,64(1):1-25.

[24] Zhang BQ, Tsiatis AA, Laber EB, et al. A Robust Method for Estimating Optimal Treatment Regimes[J]. Biometrics,2012,68(4):1010-1018.

[25] Zhao Y, Zeng D, Rush AJ, et al. Estimating Individualized Treatment Rules Using Outcome Weighted Learning[J]. Journal of the American Statistical Association,2012,107(499):1106-1118.

[26] Zhao YQ, Zeng DL, Laber EB, et al. New Statistical Learning Methods for Estimating Optimal Dynamic Treatment Regimes[J]. Journal of the American Statistical Association,2015,110(510):583-598.

[27] Zhang BQ, Tsiatis AA, Laber EB, et al. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions[J]. Biometrika,2013,100(3):681-694.

[28] Holloway ST, Laber EB, Linn KA, et al. Dyntxregime: Methods for estimating optimal dynamic treatment regimes. R package version 4.11[EB/OL]. https://CRAN.R-project.org/package=DynTxRegime.

[29] Wallace MP, Moodie E, Stephens DA. Dynamic Treatment Regimen Estimation via Regression-Based Techniques: Introducing R Package DTRreg[J]. Journal of Statistical Software,2017,80(2):1-20.

[30] Huang X, Ning J, Wahed AS. Optimization of individualized dynamic treatment regimes for recurrent diseases[J]. Statistics in Medicine,2014,33(14):2363-2378.

[31] Basu D. Randomization Analysis of Experimental Data: The Fisher Randomization Test[J]. Journal of the American Statistical Association,1980,75(371):575-582.

[32] Bang H, Robins JM. Doubly Robust Estimation in Missing Data and Cousal Inference Models[J]. Biometrics, 2005,61(4): 962-973.

[33] Gill RD, van der Laan MJ, Robins JM. Coarsening at Random: Characterizations, Conjectures, Counter-Examples//Proceedings of the First Seattle Symposium in Biostatistics:Survival Analysis[M]. New York:Springer,1997.

[34] Berkowitz RI, Wadden TA, Gehrman CA, et al. Meal Replacements in the Treatment of Adolescent Obesity: A Randomized Controlled Trial[J]. Obesity,2011,19(6):1193-1199.

[35] Laber E, Qian M, Lizotte D, et al. Statistical Inference in Dynamic Treatment Regimes[J]. Arxiv Methodology,2010.

(责任编辑:张　悦)