双马尔可夫决策过程联合模型

王蓁蓁[1] 邢汉承[2]
(南京大学计算机科学与技术系 南京 210093)[1] (东南大学计算机科学与工程学院 南京 210096)[2]

摘 要 人类在处理问题中往往分为两个层次,首先在整体上把握问题,即提出大体方案,然后再具体实施。也就是说人类就是具有多分辨率智能系统的极好例子,他能够在多个层次上从底向上泛化(即看问题角度粒度变"粗",它类似于抽象),并且又能从顶向下进行实例化(即看问题角度变"细",它类似于具体化)。由此构造了由在双层(理想空间即泛化和实际空间即实例化)上各自运行的马尔可夫决策过程组成的半马尔可夫决策过程,称之为双马尔可夫决策过程联合模型。然后讨论该联合模型的最优策略算法,最后给出一个实例说明双马尔可夫决策联合模型能够经济地节约"思想",是运算有效性和可行性的一个很好的折中。

关键词 马尔可夫决策过程,增强学习,最优策略
中图法分类号 TP18 **文献标识码** A

Associated Model of Bi-Markov Decision Processes

WANG Zhen-zhen[1] XING Han-cheng[2]
(Department of Computer Science & Technology, Nanjing University, Nanjing 210093, China)[1]
(School of Computer Science & Engineering, Southeast University, Nanjing 210096, China)[2]

Abstract Human thought is often divided two levels while dealing with problems. First people always treat problems from a whole perspective, i. e., they have a general plan, then they specifically deal with details. The human itself is a good example for having a multi-resolutional characteristic. It can not only generalize bottom-up among multi-levels (the granule of viewpoint about problem becomes "rough", analogous to abstract), but also instantiate top-down (the granule of viewpoint becomes "thin", analogous to specification). So we constructed a semi-Markov decision process consisting of two Markov decision processes running respectively on two levels—the ideal space (generalization) and the actual space (instantiation). It is called an associated bi-Markov decision model. Then we discussed how to find optimal policy under this associated model. Finally an example was given to show that the associated bi-Markov decision process model can economically economize "mind" and is a good tradeoff between the computational validity and computational feasibility.

Keywords Markov decision processes, Reinforcement learning, Optimal policy

1 引言

马尔可夫决策过程及其增强学习强有力的方法的运用,使得人们对它们的研究兴趣逐渐增长[1-12]。不幸的是,它一直受到几个方面的困扰。1. 实际问题往往庞大复杂,马尔可夫决策过程很难解决自己的计算可行性问题;2. 实际问题往往是(环境)部分可观察的或者是不可观察的,且大都具有非马尔可夫性,或者是具有半马尔可夫性。为了解决第一个问题,许多作者[2,4]提出了逻辑马尔可夫决策过程,力图用更简洁的方式表达问题,当然还有文献[8,9]利用代数来缩小状态空间的努力(文献[9]还是有关半马尔可夫空间的);为解决第二个问题,文献[1,6,7]讨论了马尔可夫性和半马尔可夫性的关系,特别是文献[1]引入了 options 概念,使得马尔可夫决策过程(MDPs)和半马尔可夫决策过程(SMDPs)之间建立了有机联系,此外文献[5]还提出用分层方法(即多 agent 方法)解决部分可观察问题以及有关半马尔可夫决策问题。这些方法和模型的提出虽然在很大程度上缓和了前面提到的两类问题,但是有时也会引起新的问题,例如在关系抽象层次上,一般决策问题不再具有马尔可夫性,参看文献[4],即出现问题之间转换现象。更有甚者,上述模型还以各自侧重点的不同而呈现不同的模样。本文作者在研究了上述各文献以后,特别受到人类思维范式的启发,提出了由两个层次上的马尔可夫决策过程组装的一个半马尔可夫决策过程联合模型,以统一上述文献的各种思想和概念,同时分享它们共同的优点。众所周知,人类思维明显分为两个层次,首先是把握全局,在抽象层次上寻找一个大概经略,做到"心中有数",然后再亲躬力行,在实际层面上具体施行自己的方略。因此用理性状态空间表达人类的抽象层次,由该空间上的马尔可夫决策过程,

到稿日期:2008-10-27 返修日期:2009-01-09 本文受国家自然科学基金(90412014,60803061),江苏省自然科学基金(BK2008293)资助。
王蓁蓁(1975-),女,博士,主要研究方向为马尔可夫决策过程、增强学习等,E-mail:wangzhenzhen@seu.edu.cn;邢汉承(1938-),男,教授,博士生导师,CCF 会员,主要研究方向为人工智能、机器学习、模式识别。

找到一系列的谋略，来逐步实现人类的各个理想，这个理想-谋略系列便是人们在解决问题时首先在心中出现的一个大概经略。用实际状态空间表达人类的实践层次，它由一些具体的马尔可夫决策过程簇所组成，簇里每个具体过程都相当于人类心中大概经略里某个方略的实施方案和途径。这两个层次上的决策过程都具有马尔可夫性，然而把它们组装在一起的联合模型却是半马尔可夫性的，这样便结合了马尔可夫性和半马尔可夫性。由于理想状态空间内涵特别丰富，它既可以表达逻辑内容，也可以表达观察内容，因此联合模型确实涵盖了许多模式。特别是理想状态空间上的理想和谋略（它相当于抽象状态空间的行动）远比实际状态和行动要少得多，这在复杂问题中更为明显，加上由联合模型寻找最优策略较简单并在平均意义上是最优的，所以联合模型很好地解决了计算复杂性问题，是一个有效性和可行性成功的折中方法。

本文第 2 节引入联合模型的一般结构，论述了两个空间上的马尔可夫性和组装的整体空间上过程的半马尔可夫性；第 3 节引入联合模型的两个重要类型，从中可以看出理想状态空间的内涵特别丰富，可以涵盖许多文献中的不同类型；第 4 节介绍最优策略的求法，从中可以看出联合模型的最优策略远比实际空间中最优策略的求法简单，但是它只在平均意义上具有最优性，即它是关于模型最优；第 5 节用一个简单实例阐述论证了上述各个概念和思想；最后是结论。

2 联合模型的一般结构

2.1 两个空间

设$(S,\mathfrak{B}(S))$为带有 σ-代数的实际状态空间，当 S 是有限集时，一般取 $\mathfrak{B}(S)$ 为 $\mathscr{P}(S)$，即由所有 S 上的子集所构成的幂集。本文只讨论 S 是有限的情况。

设$(Y,\mathfrak{B}(Y))$为带有 σ-代数的理想状态空间，同样当 Y 是有限集时，一般的 $\mathfrak{B}(Y)$ 为 $\mathscr{P}(Y)$。本文只讨论 Y 是有限的情况。

令 $\mathfrak{X}=S\times Y$，$\mathfrak{B}=\mathfrak{B}(S)\times\mathfrak{B}(Y)$，即$(\mathfrak{X},\mathfrak{B})$为带有乘积 σ-代数的乘积空间，称 $\mathfrak{X}$ 为联合状态空间，其中的点记为 $x=(s;y)$，$s\in S$，$y\in Y$。当 S,Y 都是有限集时（正如本文所考虑的情况），则 $\mathfrak{X}$ 也是有限集，而 $\mathfrak{B}$ 即为“二维”空间 $\mathfrak{X}$ 上所有子集所构成的 σ-代数。称$(Y,\mathfrak{B}(Y))$为理想状态空间或者称为理念状态空间，它是由对实际问题分析引入的空间，在其上定义一个马尔可夫决策过程，简记为 $\mathscr{M}=(Y,O,\tau,\rho)$。其中：

$Y=(y^1,y^2,\cdots,y^k)$，y^i 称为理想或者理念状态，简称理想，或者理念，$i=1,2,\cdots,k$；

$O=(o^1,o^2,\cdots,o^l)$，o^j 称为谋略，$j=1,2,\cdots,l$；

$\tau:Y\times O\times Y\to[0,1]$，$\forall y\in Y$，$\forall o\in O_y$（$O_y$ 表示 agent 已达到理想状态 y，并在该处为了今后理想，可以采取的一切谋略的集合），$\forall y'\in Y$，$\tau(y,o,y')$表示一个概率，它是 agent 实现了理想 y 并采取谋略 o 而获得新理想 y' 的概率。τ 称为理想转移概率，它满足下述关系：

$\forall y\in Y$，$\forall o\in O_y$，有 $\sum_{y'\in Y}\tau(y,o,y')=1$。

$\rho:Y\to\mathscr{R}$，$\forall y\in Y$，$\rho(y)$表示 agent 获得理想 y 后所获得的报酬。假设 $\forall y\in Y$，$\rho(y)$有界。

称$(S,\mathfrak{B}(S))$为实际状态空间，简称为状态空间。在其上定义一簇马尔可夫决策过程，即 $\forall y\in Y$，有一个马尔可夫决策过程 $M_y=(S,A,T_y,R)$相连。实际上，只有 k 个马尔可夫决策过程。对于每一个 M_y，S,A,R 都是相同的，只有转移概率可能依赖参数 y。详细描述如下：$S=\{s^1,s^2,\cdots,s^n\}$，s^i 称为实际状态，简称状态，$i=1,2,\cdots,n$；

$A=\{a^1,a^2,\cdots,a^m\}$，a^j 为 agent 可采取的实际行动，简称行动，$j=1,2,\cdots,m$；

$T_y:S\times A\times S\to[0,1]$，$\forall s\in S,a\in A_s$（$A_s$ 表示 agent 处于状态 s 处可采取的行动集合），$\forall s'\in S$，$T_y(s,a,s')$表示一个概率，它是 agent 处于状态 s，采取行动 a，从而转移到新状态 s'的概率。T_y 称为实际转移概率，简称为转移概率，它满足下述条件：

$\forall s\in S$，$\forall a\in A_s$，$\sum_{s'\in S}T_y(s,a,s')=1$。

$R:S\to\mathscr{R}$，$\forall s\in S$，$R(s)=r\in\mathscr{R}$，表示 agent 处于 s 处所获得的报酬 r。假设 $\forall s$，$R(s)$有界。

为了节省符号，也用 $s(t),y(t)$分别表示取值于 S,Y 空间上的随机过程，$t\in T$，其中 T 表示时间域，它只取离散值：0，1，2…。于是 $x(t)=(s(t);y(t))$即为取值于联合空间 $\mathfrak{X}$ 上的随机过程。因此，当 t 固定时，$y(t)$是$(Y,\mathfrak{B}(Y))$上的随机变量，假设它服从由理想转移概率 τ 所刻画的分布。而当 t 与 $y(t)$都固定时，$s(t)$即 $\mathfrak{X}(t)$的第一分量，是$(S,\mathfrak{B}(S))$的随机变量，但假设它服从由 $M_{y(t)}$ 中的实际转移概率 $T_{y(t)}$ 所刻画的分布。

2.2 两个层次

$(Y,\mathfrak{B}(Y))$“纯粹”是一个理念世界，它往往属于战略层次；$(S,\mathfrak{B}(S))$却是“真实”的现实世界，它往往属于战术层次。在人们思考问题时，两个层次上的马尔可夫决策过程完全是由谋略集合 O 所联系的，在理念世界里，每一个谋略都是不透明的，作为一个整体“暗箱”操作着。然而在现实世界里如果 agent 想获得理想 y，每一个谋略 o 在与之相连的 M_y 中被打开，由 o 的内部结构操纵 $s(t)$在 M_y 中的运动。

如上所述，对于每一个理念 y，都有一个相应的马尔可夫决策过程 M_y 相连。也就是说整体马尔可夫决策过程 $x(t)=(s(t);y(t))$，当第二分量 $y(t)=y$ 固定时，第一分量 $s(t)$是M_y 马尔可夫决策过程。如果表示 agent 运动的 $s(t)$，想获得理想 y，它就必须在 M_y 中按某个谋略 o 行事，称能够获得理想 y 的谋略 o 为与 M_y 相连的谋略，并用符号“$\mathbb{R}_y$”表示所有这样谋略 o 的集合，如此谋略在 M_y 中施予后能获得理想 y。沿用文献[1]关于 options 的类似定义[1]，假设谋略 o 由 3 个成分所组成，即 o 是一个三元组 $o=\langle I,\pi,\beta\rangle$，其中 $I\subseteq S$ 是初始状态集合，$\pi:S\times A\to[0,1]$为一个策略；$\beta\subseteq S$ 为终止状态集合。一个谋略 $o=\langle I,\pi,\beta\rangle$ 能够启动，当且仅当 agent 处于的状态 $s\in I$，当该谋略被 agent 采纳以后，agent 根据策略 π 选择行动 $\pi(s,\cdot)$，比如说以概率 $\pi(s,a)$选择行动 a 后，

[1] 文献[1]options 定义中的 β 是定义在 s^+ 上的终止条件，而不是我们这里的终止状态集合，细节参考文献[1]。不过，也可以取 β 为终止条件，使模型更一般化。虽然如此，但是我们的谋略定义与文献[1]的 options 是有本质区别的。直观上，文献[1]中的 options 是“宏行动”，它在实际状态空间上实施，与行动属于“同一级”，而我们的谋略，作为整体时，它是理想空间上的“行动”，而它在实际空间里打开时，是一个具体行动或若干行动组成的“子程序”。

状态空间根据转移概率 $T_y(s,a,s')$ 转移到 s'，如果 $s'\in\beta$，则谋略终止，即 agent 获得理想 y，否则继续，于是 agent 又根据策略 π 选择下一个行动。假设选择了 b，然后状态空间以概率 $T_y(s',b,\cdot)$ 转移到 s''，视 s'' 是否属于 β 而决定是否谋略终止，如果终止，则意味着 agent 获得了理想 y，否则继续下去，如此等等。为了确定起见，把几个有关概率以定义形式重写如下：

定义 1

$O_y=\{o\mid o=\langle I,\pi,\beta\rangle$：如果 agent 获得了理想 y，正处在状态 s 处，而 $s\in I\}$；

联$_y=\{o\mid o=\langle I,\pi,\beta\rangle$：如果 agent 正在 M_y 中运动，并利用 o 能够获得理想 $y\}$。

那么，当 agent 已获得了某个理想 y，并想获得另一个理想 y'，它应该如何运作呢？这要视 agent 获得理想 y 时终止的状态 s 而定，如果 s 恰好属于与马尔可夫决策过程 $M_y=(S,A,T_y,R)$ 相连的某谋略 o 的初始集合 I，这时 o 便可启动，从而才有可能使 agent 获得理想 y'。用符号表示上述的一段话：当 agent 获得了理想 y，当且仅当 $o\in O_y\cap$ 联$_y$ 时，agent 才可以采取谋略 o，并有可能获得理想 y'。现在讨论理想转移概率：

对 $\forall y,\forall o\in O_y$，令 $\|o\|_y=\#\{y'\mid o\in O_y$ 且 $o\in$ 联$_{y'}\}$，其中“#”表示集合中元素的个数。

所以可以定义 $\tau(y,o,y')$ 如下：

$$\tau(y,o,y')=\begin{cases}\dfrac{1}{\|o\|_y}, & o\in O_y,\text{且 } o\in\text{联}_{y'}\\ 0, & \text{否则}\end{cases}$$

显然，$\forall y\in Y,\forall o\in O_y,\sum_{y'\in Y}\tau(y,o,y')=1$。上述理想转移概率是“均匀”分布，自然可以根据具体问题而定为其他分布。至于与 y 相连的实际状态空间上的马尔可夫决策过程 M_y 中的转移概率 T_y 更要视具体问题而定，在一般情况下，也可能出现所有的 T_y 是相同的分布这种现象。

2.3 总结：联合运动的半马尔可夫性

现在，已经定义了两类相互联系的马尔可夫决策过程。

在理想状态空间$(Y,\mathfrak{B}(Y))$上：$\mathcal{M}=(Y,O,\tau,\rho)$，在其上取值的随机过程为 $y(t)$，它遵从 τ 规律运动。

在实际状态空间$(S,\mathfrak{B}(S))$上：$\forall y\in Y,M_y=(S,A,T_y,R)$。其实只有 k 个过程：$M_{y^1},M_{y^2},\cdots,M_{y^k}$。在$(S,\mathfrak{B}(S))$上取值的随机过程记为 $s(t)$，当 y 固定时，它遵从 T_y 规律运动。

两个空间上运动的随机过程，完全由谋略 o 联系。当把 o 看作暗箱操作在理想空间上时，$\mathcal{M}$ 上的 $y(\cdot)$ 才是马尔可夫的。当 y 固定时，虽然 M_y 中的 $s(\cdot)$ 遵从打开的 o 行事，但它的运动规律 T_y 仍然也是马尔可夫的。如果把两个随机过程联系在一起，决定在统一空间$(\mathfrak{X},\mathfrak{B})$上，其统一随机过程 $x(\cdot)=(s(\cdot);y(\cdot))$ 却不再有马尔可夫性了。上面几段中的叙述有点含糊，当时是为了直观的需要所致。现在把它精确化。

令 $\eta_n(n=0,1,2,\cdots)$ 为取值于 $\mathcal{R}^+=[0,\infty)$ 中的随机变量，满足关系 $\eta_0=0\leqslant\eta_1\leqslant\eta_2\cdots$。$\mathcal{M}$ 上的 $y(\cdot)$，只在诸 η_n 时刻上改变(理想)状态，在两个相邻时刻之间是常(理想)状态，精确写出来为：

$y(t)=y(\eta_n)$ 当且仅当 $\eta_n\leqslant t<\eta_{n+1}$ 时，即 $y(t)$ 的轨道(如果在 Y 空间定义比如离散拓扑)是右连续分段阶梯函数。而与相关的 $x(t)$ 的运动为：

当且仅当 $\eta_n\leqslant t<\eta_{n+1}$ 时，由于 $y(t)=y(\eta_n)$，因此 $s(t)$ 在 $M_{y(\eta_n)}$ 中运动。注意，上述的诸 η_n 以及 t 皆取离散值。

η_n 的确定是由谋略 o 确定的，每个谋略 o 的执行时间既依赖初始状态 s(它是紧接在前面一个谋略的终止状态)，又依赖终止状态 s'(它又是紧接于后另一个谋略的初始状态)，所以在联合运动 $x(t)=(s(t);y(t))$ 框架下，它是半马尔可夫性的，故在联合空间$(\mathfrak{X},\mathfrak{B})$上定义的决策过程是半马尔可夫决策过程，即把$\{\mathcal{M},M_{y^1},M_{y^2},\cdots,M_{y^k}\}$称为联合半马尔可夫决策过程，简称为联合模型。

3 联合模型的两种特殊类型

3.1 逻辑型

如果理想状态空间$(Y,\mathfrak{B}(Y))$上的 $\mathcal{M}=(Y,O,\tau,\rho)$ 是逻辑马尔可夫决策过程，则称联合模型为逻辑联合模型。

$\mathcal{M}=(Y,O,\tau,\rho)$ 的逻辑背景如下：存在一个一阶逻辑字母表Σ，除了常量和变量以外，Σ 是一个关系字母表 r(目 $m\geqslant 0$)和函数符号表 f(目 $n\geqslant 0$)的集合。如果 $m=0$，则相应的 r 为命题，如果 $n=0$，则相应的 f 为常数。为了简单起见，本文只讨论 f 为常数的情形。于是，称常数和变量为项。如果 r 是 m 目的关系符号，则称 $r(t_1,t_2,\cdots,t_m)$ 为原子，其中诸 t_i 为项。一个原子的集合则称为一个合取。如果存在一个置换 θ，使得两个合取 A,B 满足 $B\theta\subset A$，则称 B 是 θ 包含于 A。记号 hb_Σ 表示Σ的 Herbrand 基，它是由Σ构造的所有基础原子(即不包含变量的原子)的集合。

Y 中的每一个元素 y 是Σ中的一个合取，今后在本类型中也可称它为抽象状态，Y 是所有抽象状态的集合。O 中的元素也可称为抽象行动，其实它常常是抽象行动组，而理想状态转移概率 τ 也可命名为抽象转移概率，$\tau(y,o,y')$ 是如下形式表达式的缩写：$y'\xleftarrow{\tau;\rho;o}y$，它的直观语义是，agent 从抽象状态 y 采取抽象行动 o(或是抽象行动组，下同)，达到抽象状态 y'，其转移概率 $\tau(y,o,y')=\tau$，在 y' 处获得的报酬为 ρ，即 $\rho(y')=\rho$。称 y 为抽象行动的体，y' 为抽象行动的头，分别记为 $B(o)$ 和 $H(o)$(参看文献[2,3])。

在逻辑联合模型中，实际状态空间$(S,\mathfrak{B}(S))$上的任一个马尔可夫决策过程 $M_y=(S,A,T_y,R)$ 与 y 的联系如下：

$\forall y\in Y,T_y=T$，即所有 T_y 是同一个概率规律；

任意的抽象行动 o，如果存在下列形式的转移 $y'\xleftarrow{\tau;\rho;o}y$，则称 o 与 M_y 相连，即 o 是 agent 为了获得抽象状态 y' 可以在 M_y 中打开的抽象行动，并且 $o\in O_y$，即 o 是 agent 处于抽象状态 y 时可采取的抽象行动。因为所有的 M_y 都有一致结构，即 $\forall y,M_y\triangleq M=(S,A,T,R)$，所以对于任意抽象状态 y 和 y'，都有 $o\in O_y$ 和 $o\in$ 联$_y$，并且在 M 中与 o 相连的形式为：$s'\xleftarrow{a;*;\cdots;*}s$($*$ 为任意行动)从 s 出发经过若干转换到 s'，但首先采取的是行动 a，如果存在一个置换 θ，使得 $y\theta\subseteq s,o\theta=a,y'\theta\subseteq s'$ 成立。

也就是抽象行动 $o=\langle I,\pi,\beta\rangle$ 的规定如下：

$I=\{s:\forall y\in B(o),\exists\theta$，使 $y\theta\subseteq s,o\theta=a$，且 $a\in A$，并且 s 满足一定条件(视具体问题而定)$\}$，

$\beta=\{s:\forall y\in H(o),\exists\theta$ 置换使得 $y\theta\subseteq s$，并且 s 满足某些条件(视具体问题而定)$\}$。

而策略 $\pi(s,a)$一般视具体问题而定，也可以采取确定性策略，即 $\pi(s,a)=1$，$\pi(s,*)=1$ 等，也就是说如果 $s'\xleftarrow{a;*;\cdots;*}s$ 与 $y'\xleftarrow{\tau;\rho;o}y$ 在上述意义上相连，且 $s\xrightarrow{a}s_1\xrightarrow{*}s_2\xrightarrow{*}\cdots\xrightarrow{*}s'$，则对所选择的动作是确定的。

直观上，当 agent 处于抽象状态 y，相应的实际状态为 s，如果 agent 想转移到抽象状态 y'，那么它应该寻找抽象行动 o，该行动的体为 y，而头为 y'，即有形式：$y'\xleftarrow{\tau;\rho;o}y$。这时，agent 即从 s 处开始在 M 中运动，选择行动 $a\in A_s$，且存在 θ，使 $y\theta\subseteq s$，$o\theta=a$，在 M 中遵从 T 规律，如果以概率比如 $T(s,a,s_1)$转移到状态 s_1，若 $y'\theta\subseteq s_1$，则终止运行，否则根据 $\pi(s_1,\cdot)$选择行动，比如选择到 b，再以概率比如 $T(s_1,b,s_2)$转移到 s_2，若 $y'\theta\subseteq s_2$，则终止运行，即 agent 已获得抽象状态 y'，否则继续运行，如此等等。自然，由于 τ 的意义，agent 获得抽象状态 y' 也是以一定概率保证的。

由于本文只强调有关逻辑马尔可夫决策过程可以纳入本文提出的联合模型框架，因此有关逻辑马尔可夫决策过程的详细情况可以参见文献[2,10]，本文的符号也沿用于文献[2]。但是，由于联合模型分离了抽象层次和具体层次，仅仅通过谋略在具体空间的实现就能把逻辑马尔可夫决策过程和与它相应的常规马尔可夫决策过程联系起来，且谋略是在现在文献中逻辑马尔可夫过程的抽象行动的扩展（虽然沿用同样的名称指称谋略），它往往是一系列抽象行动的组合，在某种意义上，它是宏抽象行动，再加上它作为暗箱在抽象层次操作，所以联合模型中在理想空间上的 $\mathcal{M}$ 马尔可夫决策过程的操作与现存文献中逻辑马尔可夫决策过程的操作是有本质区别的，特别是在 $\mathcal{M}$ 中可以用普通的常规 MDPs 上的增强学习方法进行学习以及寻找最优策略，即它可以完全“独立”于实际空间上的马尔可夫决策过程进行操作。这就说明了联合模型不仅涵盖了逻辑马尔可夫决策过程，而且也简化了后者寻找最优策略的程序。

3.2 目标型

如果理想空间$(Y,\mathfrak{B}(Y))$是实际空间$(S,\mathfrak{B}(S))$的子空间，这时每一个理想可以看作是在$(S,\mathfrak{B}(S))$中追求的一个目标状态，故称由它们构成的联合模型为目标联合模型。

$\mathcal{M}=(Y,O,\tau,\rho)$的逻辑背景如下：正在解决的实际问题是寻找某个目标，它在实际空间$(S,\mathfrak{B}(S))$里。由于获得最终目标的途径很复杂，一时很难找到某个最优策略得到它，因此就分解总目标为各个子目标，或者说是分阶段去完成寻找任务。在 $\mathcal{M}$ 中，Y 中的每一个元素 y 都是 S 中的一个状态，它是作为（子）目标出现的，在目标联合模型中也可称它为目标状态，Y 便是所有目标状态的集合。这时谋略 o 成为实现（子）目标的手段，直观上它是宏运算子或者是子控制器，即是一个子程序。沿用文献[1]中的术语，也可称它为 options，所有 options 的集合便是 O。于是，由某个 option 引起的转换 $y'\xleftarrow{\tau;\rho;o}y$，其语义是，agent 已获得子目标 y，采取 o 行动后，以概率 τ 达到子目标 y'，如果达到子目标则获得 ρ 奖赏。简记为 $\tau(y,o,y')=\tau,\rho(y')=\rho$。

实际状态空间$(S,\mathfrak{B}(S))$上诸 MDPs 过程 $M_y=(S,A,T_y,R)$与 y 的联系是明显的。因为每个子目标 y 的实现都是依赖于与之相连的谋略 $o=\langle I,\pi,\beta\rangle$，并且 y 的实现就是某个具体状态出现，所以谋略 o 即 option 的内部结构在具体空间中打开时，可以认为 $I=\{s:s\in O_s$，即 s 满足执行 o 的条件$\}$，$\beta=\{s:s=y$ 或者 $y\subseteq s\}$，$\pi(s,a)$也可以是确定的，即 $\pi(s,a)=1$。在特殊情况下，甚至可以认为对于所有的 y，M_y 都是一致的，即为同一个马尔可夫决策过程 M。下面叙述中的 M_y 满足这里的假定：$\forall y,M_y=M$。

整体运动 $x(t)=(s(t);y(t))$可以这样刻画，当 agent 完成了某个目标 y，即 agent 处于的实际状态 s 满足条件：或者 $s=y$，或者 $y\subseteq s$，符号 $y\subseteq s$ 表示目标状态包含在状态 s 中。如果 s 满足某个(option)o 的执行条件，换言之，$o=\langle I,\pi,\beta\rangle$，$s\in I$，则当agent 决定执行 o 时，它按策略 $\pi(s,\cdot)$选择实际行动，比如是 a，然后按照 $T(s,a,\cdot)$到达某个状态 s_1，如果 $s_1\in\beta$，则 o 执行终止，agent 到达某目标比如 y'，它或者等于 s_1 或者可以导致 s_1，即或者 $y'=s_1$，或者 $y'\subseteq s_1$。如果 $s_1\notin\beta$，则 agent 又按策略 $\pi(s_1,\cdot)$选择实际行动，比如是 b，然后按照规律 $T(s_1,b,\cdot)$转移到某个状态 s_2，如果 $s_2\in\beta$，则 o 执行终止，agent 到达目标 y'，即 $s_2=y'$，或 $y'\subseteq s_2$，否则继续进行，如此等等。直观上，agent 根据具体问题和总目标，把任务分解成许多阶段，每个阶段都有一个相应的子目标，当它把这些子目标按最优策略筹划排序以后，便可以在实际中具体地一件一件完成，这便相应地在状态空间上分别寻找完成每个子目标的最优策略，从而达到最好效果。

通过这两个特殊类型的描述，可以看出本文提出的联合模型确实是符合人类思维规律和人类行为规律的广泛有效的模型，且也是一个可能有广泛运用领域的模型。

4 最优策略的求法

联合模型的最优策略的求法是很明显的，它分为两种情况。

I. 确定情况

所谓确定情况是指在联合模型中的 $\mathcal{M}$ 和诸 M_y 中的转移概率都是确定的，即无论是理想转移和实际转移中的谋略和行动都有确定的效果，且每一谋略里的策略 π 也是确定的。在这种特别情况下，寻找模型的最优方案是很简单的，它可以按以下几个步骤进行。

1）确定初始状态 s_0 和终止状态 $s_{止}$。初始状态是面临问题时的原始状态，终止状态反映问题解决时的情景，例如它可以由目标函数所确定。

2）在理想状态空间上寻找最优策略：按照传统方法在 $\mathcal{M}$ 中寻找。

3）根据初始状态 s_0，由它确定初始理想 y_0，根据终止状态 $s_{止}$，由它确定终止理想 $y_{止}$，并按 $\mathcal{M}$ 中的最优策略，寻找从 y_0 到 $y_{止}$ 的理想-谋略链：

$$y_0\xrightarrow{o_1}y_1\xrightarrow{o_2}y_2\cdots\cdots\xrightarrow{o_q}y_q=y_{止}$$

其中，o_i 是与 y_i 相连的谋略，$i=1,2,\cdots,q$。

4）相应地在 $M_{y_1},M_{y_2},\cdots,M_{y_q}$ 过程里分别寻找最优策略。

5）根据各个过程的最优策略和理想链，寻找实际状态-行动链。

s_0 为初始状态，$s_1,s_2,\cdots,s_{l_1}$ 是相应于 M_{y_1} 过程里的状态链，在 s_{l_1} 处 agent 获得了理想 y_1，并在该处出发，通过相应于 M_{y_2} 过程里的状态链 $s_{l_1},s_{l_1+1},\cdots,s_{l_2}$ 的实现，在 s_{l_2} 处获得理想 y_2，并在 s_{l_2} 处出发，直到通过相应于 M_{y_q} 过程里的状态链

$s_{l_{q-1}}, s_{l_{q-1}+1}, \cdots, s_{l_q}$ 的实现，在 s_{l_q}（$=s_{止}$）处获得最终理想 y_q，即 $y_{止}$。相应于上述每个实际状态，agent 相应采取的行动记为 $a_0, a_1, \cdots, a_{l_1}, a_{l_1+1}, \cdots, a_{l_2}, \cdots, a_{l_{q-1}}, a_{l_{q-1}+1}, \cdots, a_{l_q-1}$，具体表示之：

$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \cdots \xrightarrow{a_{l-1}} s_{l_1} \xrightarrow{a_l} s_{l_1+1} \cdots \xrightarrow{a_{l-1}} s_{l_2} \xrightarrow{a_l} s_{l_2+1} \cdots \xrightarrow{a_{l-1}} s_{l_q} = s_{止}$$

上述实际链以及相应的行动序列，只是根据模型寻找的一套最优方案，它可以作为对 agent 的建议。因为可以完全置理想状态空间于不顾，纯粹在实际空间中寻找最优方案，在某种时候，建议的上述方案可能并不比后者好。因为前面说过，理想空间里的元素个数往往大大少于实际空间里元素的数目，所以它提出的目的之一，就是缩小寻找范围，使得许多实际问题可行，故上述方案是有效性和可计算性的一个最好的折中。

II. 不确定的情况

1）在理想状态空间里，可以用传统方法寻找最优策略。

2）在诸 M_y 过程里，可以“动态”地寻找最优策略。这里动态的意思是，当 agent 运作到某一理想状态时，根据 $\mathcal{M}$ 过程中的最优策略选择某一个策略 o，那么只有在能打开 o 的具体过程 M_y 里才能寻找到实现 o 的最优策略。

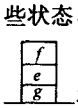
3）也可以用上述确定情况下的方法寻找到最优方案踪迹图，因为环境是不确定的，上述踪迹图只是 agent 的参考“坐标”，即在一系列关键点 $s_{l_1}, s_{l_2}, \cdots, s_{l_q}$ 上注意是否到达预定理想，如果没有，则继续在相应理想空间内寻找相应理想并在实际的 M_y 中运动即可。

前面已经说过，Y 中的元素较少，在 $\mathcal{M}$ 中用传统方法寻找最优策略是较容易的，而只在执行某个谋略 o 时，才在打开 o 的某个具体马尔可夫决策过程 M_y 中寻找最优策略是明智的，因为这时的 M_y 中的状态空间 S 虽然庞杂，但现在仅根据谋略 $o=\langle I,\pi,\beta\rangle$ 行事，许多不必要的转移和选择都不在考虑之中，正如 Russell 幽默地说过：如果你从来不准备跳下悬崖，那么你不必为跳下去的“效用”值操心。

和确定情况一样，如果纯粹在实际空间中寻找最优策略，那么所获得的效用值可能比用联合模型所获得的相应值要高，例如可以参看文献[1]、文献[2]与本模型相似情况的论述。

5 实例

我们用最简单积木世界中的逻辑联合模型来阐述上面的思想，并且为了简单起见，假定联合模型是确定情况。

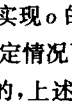
假定 e,f,g 是 3 个不同的积木块，于是实际状态空间 $(S,\mathfrak{B}(S))$ 中只有 13 个状态，其 σ-代数由 2^{13} 个子集合所组成。可以用熟知的（下面还要解释）逻辑符号 on, cl 等表示这些状态，例如用 $on(f,e)\wedge on(e,g)\wedge on(g,fl)\wedge cl(f)$ 表示  状态。注意，在如此表示法中，并没有写全逻辑关系式，例如省略了指称 e,f,g 都是积木块的合取项 $bl(e), bl(f), bl(g)$，这也仅是为了可读性的原因，以下情况类似。在 $(S,\mathfrak{B}(S))$ 空间上，agent 总共可采用 9 个行动（不包括停止行动）：e 移到 f 上，e 移到 g 上，f 移到 e 上，f 移到 g 上，g 移到 e 上，g 移到 f 上，e 移到地板上，f 移到地板上，g 移到地板上。当然，这些行动的执行是有条件的，即对于某个具体状态 s，当且仅当该行动的条件满足时，agent 才能在这个状态下采取该行动。例如对于上述状态 $on(f,e)\wedge on(e,g)\wedge on(g,fl)\wedge cl(f)$，agent 只能执行“把 f 移到地板上”行动，用符号表示就是 $mv(f,fl)$。

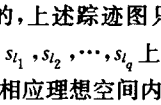
假设现在 agent 的任务，是把这 3 个积木块堆砌成 $on(e,f)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e)$ 式样。为了寻找最优策略，可以在 $(S,\mathfrak{B}(S))$ 上引入一个马尔可夫决策过程 $M=(S,A,T,R)$，其中 S 即为上述的 $(S,\mathfrak{B}(S))$ 中的状态空间，A 为上述的所有行动的集合，假定所有行动都是确定的，则 $T(s,a,s')=1$，其中 $s,s'\in S, a$ 为某行动且 $a\in A$。对于 R，例如不妨设 $\forall s\in S$，如果 s 不是目标状态，$R(s)=-1$，否则 $R(s)=1$。

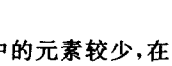
现在考虑理想状态空间 $(Y,\mathfrak{B}(Y))$。

字母表 $\sum=\{on,mv,cl,e,f,g,fl,E,F,G\}$。小写字母表示常量，$e,f,g$ 为 3 个具体的木块，fl 表示地板，大写字母表示积木块变量。谓词 on, mv 是二目关系，$on(E,F)$ 表示积木 E 在积木 F 上；$mv(E,F), mv(E,fl)$ 分别表示把 E 放到 F 上，E 放到地板上，cl 表示一目关系，$cl(E)$ 表示相应的 E 上没有其他木块。为了简洁，省略了很多谓词，例如指称 E 是积木块的谓词 $bl(E)$。

除了吸收状态以外，只有 3 个可以识别的抽象状态，用逻辑式表示如下：

Y_1:  $on(E,F)\wedge on(F,G)\wedge on(G,fl)\wedge cl(E)$

Y_2:  $on(E,F)\wedge on(F,fl)\wedge on(G,fl)\wedge cl(E)\wedge cl(G)$

Y_3:  $on(E,fl)\wedge on(F,fl)\wedge on(G,fl)\wedge cl(E)\wedge cl(F)\wedge cl(G)$

这 3 个抽象状态组成理想状态空间 $(Y,\mathfrak{B}(Y))$，Y 只有 3 个元素，$\mathfrak{B}(Y)$ 只有 $2^3=8$ 个子集，它们显然大大少于实际空间相应数目。

除了停止运动以外可以考虑的谋略只有两个：$o_1=\langle I,\pi,\beta\rangle=\{mv(E,fl)\}$，其中 $I=\{s: s\neq on(e,f)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e); s\text{‘}cl(E)\}$。关于策略 π，用 $\pi(s)$ 表示 agent 处于 s 时的选择，$\pi(s,a)$ 表示选择行动 a 的概率，则

$$\pi(s)=\begin{cases} mv(E,fl), & \text{若 } s\text{‘}cl(E)\wedge\neg on(E,fl) \\ \text{按 } M \text{ 中的转移规律选择行动}, & \text{否则}\end{cases}$$

但无论何种情况，对于可选的行动都以 $\pi(s,a)=1$ 选择之。$\beta=\{s: s=on(e,fl)\wedge on(f,fl)\wedge on(g,fl)\wedge cl(e)\wedge cl(f)\wedge cl(g)\}$，注意 $s\text{‘}cl(E)$ 表示状态 s 蕴涵 $cl(E)$ 成立，$\neg$ 表示逻辑非。直观上，o_1 表示 agent 准备把所有对象都放置到地板重新组合意图，即 o_1 是一组行动，该组行动导致的结果是所有积木块都放置在地板上。

$o_2=\langle I,\pi,\beta\rangle=\{mv(E,F)\}$，其中 $I=\{s: s=on(E,fl)\wedge on(F,fl)\wedge on(G,fl)\wedge cl(E)\wedge cl(F)\wedge cl(G)\}$，关于策略 π，规定

$$\pi(s)=\begin{cases} mv(E,F), & \text{若 } s\text{‘}cl(E)\wedge cl(F) \\ \text{按 } M \text{ 中的转移规律选择行动}, & \text{否则}\end{cases}$$

但无论何种情况，对于可选行动都以 $\pi(s,a)=1$ 选择之。$\beta=\{s: s=on(e,f)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e)\}$。直观上，$o_2$ 表示 agent 准备把所有对象顺次组合成堆的意图，即 o_2 是一组行动，该组行动导致的结果是所有积木块按要求顺序成

堆。谋略(即抽象行动组或抽象行动序列)的个数与实际行动相比也大大减少了,因此在理想状态空间构造的马尔可夫决策过程 $\mathcal{M}=(Y,O,\tau,\rho)$ 上寻找最优策略就方便得多了。具体到本例,由于 $Y=\{y_1,y_2,y_3\}$,$O=\{o_1,o_2\}$,假如约定所有的谋略都是确定的,并且在吸收状态 absorb 上规定 $\rho(absorb)=1$,而对于其他的理想,规定 $\rho(y)=-1$,则有如下转移概率,即得到 τ 的定义(附加一个停止运动定义):

$$y_3 \xleftarrow{1:-1:o_1} y_1,\ y_3 \xleftarrow{1:-1:o_1} y_2,\ absorb \xleftarrow{1:1:o_2} y_3,$$

$$absorb \xleftarrow{stop} absorb$$

最后,讨论最优策略的求法。

1)确定初始状态。假定 $s_0=on(f,g)\wedge on(g,fl)\wedge on(e,fl)\wedge cl(f)\wedge cl(e)$,终止状态是由任务确定的,它早已规定为 $s_{止}=on(e,f)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e)$。

2)在 $\mathcal{M}=(Y,O,\tau,\rho)$ 中寻找最优策略 $\pi_{\mathcal{M}}$,显然 $\pi_{\mathcal{M}}(y_1)=o_1$,$\pi_{\mathcal{M}}(y_2)=o_1$,$\pi_{\mathcal{M}}(y_3)=o_2$。该最优策略就是著名的 unstack-stack 策略(在规划领域内)。

3)根据 s_0 和 $s_{止}$,以及上面规定的理想的转移概率,能得到理想(抽象状态-谋略)链为:

$$y_2 \xrightarrow{o_1} y_3 \xrightarrow{o_2} absorb$$

4)在相应的 $M_y(=M)$ 过程里寻找最优策略。为了获得 y_3,显然从状态 $s_0=on(f,g)\wedge on(g,fl)\wedge on(e,fl)\wedge cl(f)\wedge cl(e)$,用行动 $a_1\triangleq mv(f,fl)$ 一步即可达到,它当然就是为了获得 y_3 在实际空间中的最优策略。为了获得 absorb 状态,在 M 中从上述理想的终止状态 $on(e,fl)\wedge on(f,fl)\wedge on(g,fl)\wedge cl(e)\wedge cl(f)\wedge cl(g)$ 用两步实际转移达到,即先用行动 $a_2\triangleq mv(f,g)$ 到达状态 $on(e,fl)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e)\wedge cl(f)$ 再用行动 $a_3\triangleq mv(e,f)$ 到达吸收状态 $on(e,f)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e)$。显然它是从 $on(e,fl)\wedge on(f,fl)\wedge on(g,fl)\wedge cl(e)\wedge cl(f)\wedge cl(g)$ 到达 $on(e,f)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e)$ 最优策略,因为它获得的总报酬为 0。假如 agent 第一步用行动比如 $mv(f,e)$,则出现状态 $on(f,e)\wedge on(e,fl)\wedge on(g,fl)\wedge cl(f)\wedge cl(g)$,因为它不是终止状态,所以 agent 为了实现最终理想,在按谋略 o_2 的行事时,必须继续"盲目"行走,即不停地运用 M 中的转移运动,直到出现吸收状态为止。虽然,这是我们在前面叙述谋略意义时所规定的合法运动,但类似这样运动所获得的报酬显然是负数,所以 a_2,a_3 相继运动形成为了获得第二个理想(即 absorb)的最优策略。

5)实际状态-行动链:

$$on(f,g)\wedge on(g,fl)\wedge on(e,fl)\wedge cl(e)\wedge cl(f)\xrightarrow{mv(f,fl)} on(e,fl)\wedge on(f,fl)\wedge on(g,fl)\wedge cl(e)\wedge cl(f)\wedge cl(g)\xrightarrow{mv(f,g)} on(f,g)\wedge on(g,fl)\wedge on(e,fl)\wedge cl(f)\wedge cl(e)\longrightarrow on(e,f)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e)。$$

注意,正如前面强调过的,用联合模型寻找到的最优途径往往并不比单纯在实际空间中寻找到的最优途径好,本例更尖锐地指出了这个事实,agent 刻板执行一般理想操作优化程序,作了无用功 $a_1=mv(f,fl)$ 和 $a_2=mv(f,g)$,它犯了"教条主义"的错误。然而,首先注意到 Y 和 O 比 S 和 A 的个数要少得多,对于复杂问题更是如此,所以用联合模型往往更使问题实际可行,更何况用它获得的最优策略在平均意义上是很好的。其次还可以注意到,如果修改谋略 $o_2=\langle I,\pi,\beta\rangle=\{mv(E,F)\}$ 中的 I,即重新定义 $I=\{s;s\neq on(e,f)\wedge on(f,g)\wedge on(g,fl)\wedge cl(e),s\text{`}cl(E)\wedge cl(F)\}$,策略 π 和终止集合 β 不变,那么从初始状态 $on(f,g)\wedge on(g,fl)\wedge on(e,fl)\wedge cl(e)\wedge cl(f)$ 开始,就可以直接运用策略 o_2,从而在 M 中直接找到最优策略 $mv(e,f)$,即一步登天,获得最优理想。因此适当定义理想状态空间和谋略,也可以改善效率,就像刚才所证实的那样,运用联合模型可以作得和实际模型一样好。

结束语 本文提出的联合模型,把两个层次上的马尔可夫决策过程有机地组合在一起,至少有两个重要功效:1)在某种意义上,它是马尔可夫和半马尔可夫决策的统一,也是"运算"有效性和可行性的最出色的一种折中;2)它更像是一种人类思维行为,对于复杂问题,人类首先是高屋建瓴,从战略上解决问题,然后再回到实际问题,从战术上具体实现战略目标。本文作者力图开辟一条和人类既不同又相似的智力系统的研究途径,联合模型就是这种努力的一个体现。

参 考 文 献

[1] Sutton R S, Precup D, Singh S. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning [J]. Artificial Intelligence, 1999, 112: 181-211

[2] Kersting K, Raedt L D. Logical markov decision programs[C]// Working Notes of the IJCAI-2003 Workshop on Learning Statistical Models from Relational Data (SRL-03). Acapulco, Mexico, 2003: 63-70

[3] Kersting K, Raedt L D. Logical Markov Decision Programs and the Convergence of Logic TD(λ)[C]// Proceeding of The 14th International Conference of Inductive Logic Programming. Porto, Portugal, 2004: 180-197

[4] Otterlo M V. Reinforcement Learning for Relational MDPs[C]// Proceedings of the Annual Machine Learning Conference of Belgium and the Netherlands. Brussels, Belgium, 2004: 138-145

[5] Wiering M, Schmidhuber J. HQ-Learning: Discovering Markovian Subgoals for Non-Markovian Reinforcement learning[R]. IDSIA-95-96

[6] Whitehead S D, Lin L-J. Reinforcement learning of non-Markov decision processes[J]. Artificial Intelligence, 1995, 73: 271-306

[7] Das T K, Gosavi A, Marchalleck S M N. Solving Semi-Markov Decision Problems using Average Reward Reinforcement learning[J]. Management Science, 1999, 4(45): 560-574

[8] Ravindran B, Barto A G. Symmetries and Model Minimization in Markov Decision Processes[R]. UM-CS-2001-043

[9] Ravindran B, Barto A G. SMDP Homomorphisms: An Algebraic Approach to Abstraction in Semi-Markov Decision Processes[C]// Proceeding of the Eighteenth International Joint Conference on Artificial Intelligence. 2003

[10] Kersting K, Otterlo M V, Raedt L D. Bellman goes Relational [C]// Proceedings of the 21st International Conference on Machine learning. Banff, Canada, 2004

[11] 王蓁蓁,邢汉承,张志政,等. 马尔可夫决策过程两种抽象模式[J]. 计算机科学,2008,35(10):6-14

[12] 唐亮贵,刘波,唐灿,等. 基于神经网络的 Agent 增强学习模型[J]. 计算机科学,2007,34(11):156-159