文章编号:1001-2486(2006)06-0081-05

用户偏好提取 MDP 建模研究*

黄海清[1],张 平[1],张曦文[2]
(1. 北京邮电大学 电信工程学院,北京 100876; 2. 航天部第二研究院 中心军代室,北京 100854)

摘 要:将马尔可夫判决过程和智能强化学习算法相结合,给出了异构无线网络环境下用户业务偏好评估模型的技术框架。为动态环境下用户需求的感知、量化和适配特征的研究提供了基本的数学描述,对解决用户体验的评价问题和业务与业务环境的适配问题提供了新的研究思路。仿真结果表明所构建的 MDP 模型能够在多状态条件下学习用户偏好,根据用户需求智能选择业务。

关键词:效用理论;用户偏好;马尔可夫判决过程;强化学习
中图分类号:TN91 文献标识码:B

Modeling of User Preference Based on MDP

HUANG Hai-Qing[1], ZHANG Ping[1], ZHANG Xi-wen[2]
(1. School of Telecommunication Engineering, Beijing Univ. of Posts and Telecommunications, Beijing 100876, China;
2. The 2th Institute of China Aerospace Science & Industry, Beijing 100854, China)

Abstract: A technical architecture for user preference model is presented, and the nature of the problem represented within a Markov Decision Process(MDP) combined with adaptive reinforcement learning algorithm is displayed. We provided a possible candidate solution for user modeling dynamically to satisfy the user's expected preference based on minimal or missing information. It is also a exploration for the evaluation of the user experience when selecting service providers. Simulations of the user models show that the MDP model is effective for learning the user preference with multi-state profiles.

Key words: utility theory; user preference; Markov decision process; reinforcement learning

未来用户将处在一个复杂的通讯网络环境中,业务的内容和提供方式将面临巨大的变化。各种业务提供商和运营商将为用户提供丰富多彩的选择,服务的价格、速率、QoS(Quality of Service)等都会各有不同,但一切服务都将以用户的需求为中心,以为用户提供最佳的服务体验为目标。通过开发智能代理系统,对用户的需求进行建模,可以协助用户选择和协商各类无线业务。用户和代理之间可以构成一个高度耦合的自适应反馈系统,系统的判决机制必须能够快速正确地响应用户复杂的需求内容。用户偏好的复杂性,要求代理的判决机制必须能够在不确定或不完整的信息状态下作出判决。无线领域信息不完整的主要原因包括:在不同的业务或需求内容条件下,用户的偏好会有所不同;在用户偏好信息缺乏的条件下,经过组合的信息提取空间依然很大;网络本身固有特性的变化会导致业务协商双方的不确定性,如 QoS 的变化会导致业务协商双方的承诺发生变化。这种综合的复杂性及信息的缺失,使我们无法使用传统的效用分析技术如 Conjoint 算法等静态的、离线的偏好提取方案,而动态的、在线的、增量迭代式的判决机制能够在最小信息条件下自适应地改进用户模型。通过对 Chajewska 数学模型[1]的扩展,Brutilier[2]创新性地提出用 POMDP(Partially Observable Markov Process)的方法来研究动态环境下,基于不完整信息的判决问题,这为本文用户偏好评估模型的构建提供了有力的数学依据。通过将 MDP(Markov Decision Process)建模方法与智能强化学习算法 RL(Reinforcement Learning)相结合,文中给出了异构无线网络环境下移动用户的业务需求和偏好评估模型的基本技术框架,并针对模型的多状态属性进行了性能评估。

* 收稿日期:2006-06-25
基金项目:国家 863 高技术资助项目(2003AA12331004)
作者简介:黄海清(1971—),男,博士生。

1 用户偏好提取技术

偏好提取的目标是构建精确的用户模型,从而使判决系统可以通过用户模型协助用户完成任务。偏好提取过程的设计一般是为特定的判决支持框架提供必需的数据。构成这些模型的理论基础主要基于判决和效用理论[3]。而判决和多属性效用理论主要是对判决问题或场景的输出和选择作出评估。

输出由一系列属性变量的值来定义,$X=\{X_1,\cdots,X_n\}$,属性变量可以是离散的,$x_i\in\{x_{i1},\cdots,x_{im}\}$,也可以是连续的 $x_i\in[a,b]$。判决问题的输出集合 O 包含于输出空间 Ω 中,$O\in\Omega$,$\Omega=\{X_1\cdot X_2\cdots\cdots X_N\}$,可知,输出集合 O 相当大,即使是离散属性的值组合起来也很大。通常的情况下,Ω 中包含的很多输出对于当前问题都是不可实行的。为了基于输出空间 O 作出判决,判决系统经常要根据用户偏好决定输出的次序,称为偏好关联,用符号$\succeq$表示。假定 $o_i,o_j\in O$,如果 $o_i\succeq o_j$,表明 o_i,o_j 相比,用户更偏好于 o_i。偏好关联一般是由值函数推导而来的,$v(o):O\to R$。假定一个值函数 v,其推导的偏好关联为$\succeq$,则有

$$\forall\, o_a,o_b\in O,o_a\succeq o_b\Leftrightarrow v(o_a)\geqslant v(o_b)$$

值函数能够在输出集和属性变量上进行运算,有

$$\text{任意 } a,b\in X_i,a\succeq b\Leftrightarrow v(a)\geqslant v(b)$$

值函数反映了用户的值对某个输出的依赖程度。然而在许多判决场景中存在一定程度的不确定性。假定某项动作会以概率 p_1 得到输出 o_1,以概率 p_2 得到输出 o_2。在不确定的情况下,单靠值函数本身不足以作出合适的判决。因为特定的动作不再承诺特定的输出,所以需要一个更复杂的函数来评估一个判决的效用。

效用理论提供的技术能够将用户对待风险的态度结合起来考虑[3],它的主要贡献就是证明了效用函数的存在,效用函数 $u(x):O\to R$ 能够在输出空间 O 上推导出偏好关联$\succeq$,但一个偏好关联并不能推出效用函数,因为效用函数必须考虑用户对风险的态度,因此效用函数在输出空间上推导出的偏好关联是基于概率分布的,如果 P_{ri},P_{rj} 分别对应动作 a_i,a_j,则有

$$a_i\succeq a_j\Leftrightarrow\sum_{o\in O}P_{ri}(o)\,u(o)\geqslant\sum_{o\in O}P_{rj}(o)\,u(o)$$

上述关联式暗示着系统实际上在寻求具有最大期望效用(MEU,Maximum Expected Utility)的动作。在假定用户是理性的条件下,效用函数对提取系统是相当重要的。在某些情况下,一群用户可能只有一个效用函数,也可能每个用户都需要一个唯一的效用函数。由于在系统交互的初始阶段,用户的效用函数经常是未知的,构建效用函数成为很多判决系统的首要目标。实际中对用户偏好的完整描述是很难得到的,因此偏好提取的主要目标就是构建精确的效用函数和用户偏好关联表示。

2 基于 MDP 的用户偏好模型框架

用 MDP 建模框架可以比较准确地描述上述的用户建模问题。离散时间 MDP 模型[8-9]应由五重组组成:$\{S,A(i),p_{ij}(a),r(i,a),V,i,j\in S,a\in A(i)\}$。各元素分析如下:

(1) S 是系统所有可能的状态所组成的非空的状态集。通过对业务场景分析可知,偏好提取系统状态包含的主要因素有:用户状态集合、业务特征集合及用户偏好集合。其中用户状态集合用 C 表示,元素 $c\in C$ 为有限维的随机序列,包含如用户当前的位置、服务请求期限、应用等内容。集合 C 中的元素根据用户的不同目标分成各个子集,用 c^g 表示。所有可能的业务特征集合用 P 表示,集合中的元素 P 由 n 个具体特征 f_i 构成,$P=(f_1,\cdots,f_n)$,具体的业务特征会随着时间、地点、服务种类及用户漫游的状态而有所不同。用户的偏好可以用集合 U 表示,其中的元素 U 代表了用户在 c^g 和业务特征 P 的条件下,基于一定概率分布对业务的选择排序。综上分析,在某个地点(用 loc 表示),t 时刻,系统状态变量 $S\in S$ 可以表示为

$$S^t=(c^g,t,loc,P,U)\tag{1}$$

(2)MDP 的另一个重要元素 $A(i)$是在状态 i 处可用的有限决策集。它应由用户和代理共同实现,使得系统状态发生转移。在 MDP 过程中,这种状态的转移表示了系统状态的配置发生了变化。在本文讨论的问题中,用户决策的变化体现在用户位置的改变、对服务质量或价格的要求、目标应用的改变、用户偏好的改变等因素。针对某个用户 u,可以简化表示为 $A^u = \{\Delta loc, \Delta app, \Delta U, \varphi\}$,分别代表位置、应用、偏好的变化,$\varphi$ 表示当前状态下无动作。

(3)当系统在决策点时刻 n 处于状态 i,采取决策 $a \in A(i)$时,系统在下一个决策点 $n+1$ 时处于状态 j 的概率为 $p_{ij}(a)$,与决策时刻 n 无关。

(4)当系统在决策时刻点 n 处于状态 i,采取决策 $a \in A(i)$时,系统在本阶段获得的报酬函数为 $r(i,a)$。

(5) V 为获得最佳报酬设定的准则函数。

在本模型中,为了表示在当前系统状态和用户的偏好条件下用户对服务的满意程度,定义下列函数

$$f^{cg}(P_i) = \sum_{i=1}^{n} w(i,j,c^g)\, v(P_{ij}) \tag{2}$$

来进行量化分析,其中 w 是加权系数, $v(P_{ij})$是针对业务的特征进行评价的函数,函数形式应根据具体情况有所调整。

MDP 问题的一个重要特征是系统通过定义策略序列 $\pi = (\pi_0, \pi_1, \cdots)$,当系统在时刻 n 时的历史 $h_n = (i_0, a_0, i_1, a_1, \cdots, i_{n-1}, a_{n-1}, i_n)$ $(n \geqslant 0)$时($i_k \in S, a_k \in A(i_k)$分别表示系统在第 k 个时刻点所处的状态和采取的决策),策略则按 $A(i_n)$上的概率分布 $\pi_n(\cdot \mid h_n)$采取决策。对应策略 π,在 MDP 中与之相关的随机序列 $R = (R_0, R_1, R_2, \cdots)$为报酬过程,其中 $R_n = r(S_n, a_n)$是系统在时刻 n 采取决策 a_n 时获得的报酬。针对本文讨论的问题最优策略[4]的选择应保证在有限维空间内获得最佳的期望总报酬[4-5],即

$$V_N(\pi, i) = \sum_{n=0}^{N} E\{r(i_n, a_n)\}, \quad i \in S \tag{3}$$

式中的报酬 r 就是代理收到的用户对所选业务的反馈,所以最佳策略的选择就是要体现用户对所选业务的最大满意度,策略 π 又可由下式表示:

$$\pi = \arg\max E\{\sum_{n=0}^{N} f_n\} \tag{4}$$

从(3)、(4)式的分析可知,当系统越来越复杂时,难以获得用户精确模型,且动态特性为时变的时候,常规的控制方法难以解决问题。对用户偏好的学习强调的就是对变化环境的适应,强化学习方法 RL 应该是可选的重要方法之一。通过 RL 方法能够较好解决移动代理和随机环境交互获取信息后,如何获得最优策略的问题。

3 模型多状态属性分析

在我们设计的多状态模型中,获得奖赏的值将同时基于代理的动作和业务特征文件,且允许对代理的意图进行建模,在业务特征空间上来学习用户的效用函数。因此在构建模型的过程中,必须要考虑多个特征元素之间的平衡关系[6]。为简化分析,假设业务特征文件具有两个特征的向量(m,n),m,n 具体含义可以根据实际情况来定(如成本、带宽、QoS 等)。

在多状态模型中,当前业务的特征文件属性在每次代理采取相关动作之后可能会发生改变,由于文中 MDP 模型传输函数是判决性的,并假定只要条件满足,代理总会得到它需要的业务特征文件。例如 如果在 t 时刻状态 S 为(m_t, n_t),代理选择动作 X,状态$(m_{t-1}, n_t) \in P$,则 S_{t+1}:(m_{t-1}, n_t)。如果需要的业务特征文件不在 P 中,则状态保持不变。通过在固定集合中限制代理动作可以降低代理的复杂性,但代理仍旧可以在业务集合内进行探索。假定在每个步骤,代理根据用户的动作 A^u 获得一个奖赏 r,用户动作的概率分布 $p(S_i)$是基于当前业务特征文件和状态 S_i 的效用分布。为简化分析,我们将用户的效用属性用线性函数 $U(m,n) = W_m m + W_n n$ 表示,来反映对 m,n 两个特征元素的平衡关系。

设计这个效用函数是为了便于计算,但并不影响多状态模型在代理与环境之间捕获用户效用和用户动作之间的关系。随着用户模型研究的深入,设计效用函数可以采取更复杂的方式,允许对用户偏好进行更宽范围的描述。

由于模型所具有的多状态属性,用于解决单状态 BANDIT[5] 类问题的方法将不能精确学习最佳的代理动作。BANDIT 类问题的解决方法通过学习可以知道在任何状态下哪个动作产生最大奖赏,但为了使系统最大化所有奖赏值之和,代理必须考虑其他状态的值。这类问题可能的解决方法包括 DYNAMIC PROGRAM,MONTE CARLO,TD LEARNING[5-6]。其中 DYNAMIC PROGRAM 要求具有完美的 MDP 模型条件,且由于计算量过大,使得典型的 DYNAMIC PROGRAM 算法在 RL 中的应用是受限的。但 DYNAMIC PROGRAM 算法在理论上的作用仍旧很重要,是其他算法如 MONTE CARLO 或 TD LEARNING 的基础,但 MONTE CARLO,TD LEARNING 要求更少的计算量,且不对环境模型有苛刻的要求。MONTE CARLO 并不假定必备所有关于环境的完备知识,仅仅要求一些体验、状态、动作、奖赏的采样序列,但其策略学习模式是离线式的,而本文 MDP 模型的非插曲式特性需要应用在线的解决方案,且希望通过与用户的交互来学习策略。与其他两个方法相比,TD LEARNING 更适用于非插曲式的、奖赏未知的任务,它不需要完备的环境模型及系统下一状态的概率分布模型,且具备在线的、迭代的运算风格。另外 TD 是从状态的转换中进行学习,而不必等到某个插曲的结束,这使得 TD 方法更适用于我们的模型要求。

在 TD(λ)方法中,我们选择 1-STEP BACKUP 方法作为初始方法。1-STEP BACKUP TD 方法包括 SARSA,Q-LEARNING,ACTOR-CRITIC 等[5-7],Q-LEARNING 是个离线式的学习方法,它总是学习最佳策略的估值,但并不考虑使用过的策略。与 Q-LEARNING 相比,SARSA 是个在线式的学习方法,在进行动作值估计时,考虑当前采取的策略,并根据下列准则更新[5]

$$Q(S_t,a_t)\leftarrow(1-\alpha)Q(S_t,a_t)+\alpha[r_{t+1}+\rho Q(S_{t+1},a_{t+1})] \tag{5}$$

其中 S_t 是当前状态,S_{t+1}为下一状态。a_t 是当前代理动作,a_{t+1}是根据策略的下一代理动作。α 是常数加权,r_{t+1}是在下一个步骤接收到的奖赏,ρ 是折扣因子。

4 仿真结果

为了验证上文 MDP 模型在多状态情况下对用户偏好的学习能力,分别在 ε-greedy、Gibbs softmax 策略下对 SARSA 的学习能力进行了评估。其中 softmax 方法中 $T_0=10$,$u=0.01$,ε-greedy 方法中 $\varepsilon=0.01$,$\alpha=0.2$。在仿真环境中共执行10 000次任务,每个任务包含 1000 次的动作学习,奖赏的初始值是 -1 和 $+1$ 间的任意值,同时假定业务特征向量(m,n)是以(5,5)为中心,半径 R 分别为 5 和 3 的两个空间。图 1、2 分别给出了模型在不同策略、不同动态业务特征范围条件下得到的平均奖赏值。由仿真结果可知 SARSA 依据 ε-greedy 策略在多状态条件下对用户偏好的学习能力要优于 softmax 策略,同时 SARSA 在状态空间增大的情况下学习能力并未减弱,同样变得有所增强。

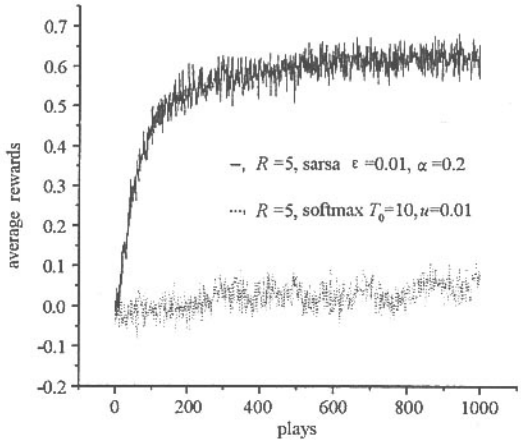

图 1 多状态用户模型性能评估($R=5$)

Fig. 1 Performance analysis of multi-state model ($R=5$)

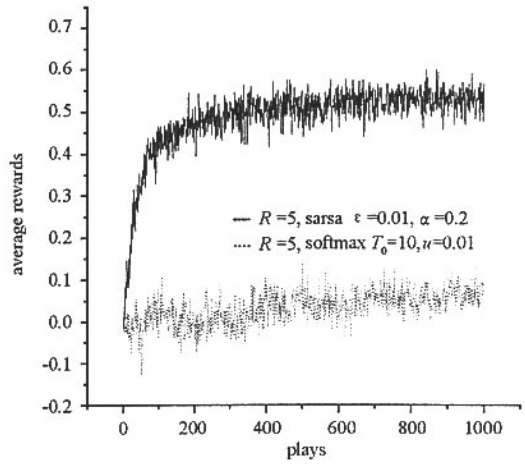

图 2 多状态用户模型性能评估($R=3$)

Fig. 2 Performance analysis of multi-state model ($R=3$)

5 结 论

本文将 MDP 建模方法和 RL 技术相结合,给出了无线异构网络环境下用户业务偏好评估模型的基本技术框架,为量化、感知用户需求和智能业务的选择与适配问题提供了新的研究思路。为进一步提高模型对用户偏好的学习能力,一方面要深入研究更加复杂 MDP 模型下智能学习算法,还要将目前采用的单一代理判决机制扩展为多代理系统的协商机制,多代理系统协商机制的研究将为用户偏好提取建模开拓一个新领域。

参 考 文 献:

[1] Chajewska U, Koller D, Parr R. Making Rational Decisions during Adaptive Utility Elicitation[A]. In Proceedings of the Seventeenth National Conference on Artificial Intelligence[C], Austin, TX, 2000:363-369.

[2] Boutilier C. A POMDP Formulation of Preference Elicitation Problems [A]. In Proceedings of American Association of Artificial Intelligence[C], Edmonton, Alberta, Canada, 2002:239-640.

[3] French S. Decision Theory: An Introduction to the Mathematics of Rationality[M]. New York, USA Halsted Press, 1986.

[4] Kaelbling L P, Moore A W. Reinforcement Learning: A Survey[J]. Journal of Artificial Intelligence Research, 1996, 4:237-285.

[5] Sutton R S, Barto A G. Reinforcement Learning[M]. MIT Press, Cambridge, MA, 1998.

[6] Boulet D P, Fraser N M. Improving Preference Elicitation for Decision Support Systems[J]. IEEE, 1995:1574-1579.

[7] Ha V, Haddawy P. A Hybrid Approach to Reasoning with Partial Preference Models[A]. In Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence[C], 1999:263-270.

[8] 胡奇英,刘建庸. 马尔可夫决策过程引论[M]. 陕西:西安电子科技大学出版社,2000.

[9] 刘克. 实用马尔可夫决策过程[M]. 北京:清华大学出版社,2004.