多代理模糊收益及策略学习

张化祥[1]　黄上腾[2]

(山东师范大学信息管理学院　济南 250014)[1]　(上海交通大学计算机科学与工程系　上海 200030)[2]

摘　要　本文研究了基于模糊知识的多代理决策问题。通过建立代理决策目标的模糊知识，我们给出了基于模糊收益的多代理决策模型，并研究了基于梯度的代理策略学习算法。

关键词　模糊集合，对策，梯度学习

Fuzzy Reward and Policy Learning in Multi-Agent Systems

ZHANG Hua-Xiang[1]　HUANG Shang-Teng[2]

(School of Information Management, Shandong Normal University, Jinan 250015)[1]

(Department of Computer Science and Engineering, Shanghai Jiaotong University, Shanghai200030)[2]

Abstract　The multi-agent decision based on fuzzy knowledge is discussed. The agent's fuzzy reward is proposed under the fuzzy knowledge of different decision goals, and a gradient learning algorithm is described to learn the agent's action policy under fuzzy reward.

Keywords　Fuzzy set, Game, Gradient learning

1　引言

经典对策论基于代理的个体理性假设，寻求策略的 Nash 均衡点，没有考虑代理的意图、信念及愿望等对决策的影响。Rao 等[1]提出的 BDI 模型给出了描述代理信念、愿望及意图的形式化框架。当代理无法准确描述其信念、愿望及意图时，可以借助于模糊集合理论，建立代理的模糊 BDI 推理模型。实际上，代理交互中存在很多不确定因素，需要建立新的代理对策机制，并在对策中充分考虑不确定因素对策略的影响。

模糊集合理论[2]研究基于模糊知识的推理。代理决策受模糊知识的影响，需要研究模糊收益下代理的决策及策略学习。目前，已经有很多成功的学习算法[4,9~11]解决确定性收益下代理的决策及学习，而代理模糊收益及策略学习的研究目前还不多见。

传统对策论假设代理对策略集合中的每个策略具有相同偏好。实际上，代理往往对不同策略赋予不同的偏好，且该偏好具有不确定性，它由代理的决策意图或信念决定。因此，需要研究意图及信念对决策的影响，并对代理的策略偏好加以描述。为更好地反映代理决策的不确定性，我们将模糊集合理论应用到多代理对策。

2　代理模糊决策模型

多代理策略组合是系统中所有代理的策略组合。决策意图不同时，代理对策略组合的偏好不同，需要建立策略组合的偏好函数。为简化问题，我们只讨论决策意图对代理策略的影响。

代理的行为不能简单地用自利(只考虑自身利益)或合作(同时考虑自身和其它代理的利益)心态描述，它是代理多种心态的组合。即代理交互中，同时存在自利、合作及其它成分。我们用模糊收益函数表示代理在不同心态下的决策利益，并由此研究模糊收益下代理策略的学习。首先，我们建立不同心态及意图下的代理策略组合隶属度函数。

隶属度在模糊集合中表示一个元素属于某个集合的程度。如给定元素 x 属于集合 A，传统集合表示为 $x\in A$。而在模糊集合中，我们以 $\mu_A(x)$ 表示 x 属于集合 A 的程度，记为 $\mu_A(x)/x\in A$。其中 $\mu_A(x)\in[0,1]$，A 为模糊集合。

设有 m 个代理，且第 i 个代理的策略集合为 Π_i，代理行动策略组合集合为 $\Pi=\{(\pi_1,\cdots,\pi_m),\forall\pi_i\in\Pi_i,i=1,\cdots,m\}$。对于 $\forall\pi\in\Pi$，代理 i 对 π 的模糊满意度评估记为 $\mu_\Pi(\pi)$。此时，建立基于论域 Π 的模糊集合：

$$A(\Pi)=\{(\pi_k,\mu_\Pi(\pi_k)),\pi_k\in\Pi,k=1,2,\cdots\} \quad (1)$$

π 的隶属度函数以 $\mu_\Pi(\pi)$ 表示，为策略组合 π 出现的可能性。

代理决策意图受心态影响。我们改进式(1)所示的模糊集合，定义同心态、不同决策意图对应的策略组合隶属度函数：

$$A'(\Pi|g)=\{(\pi_k,\mu_\Pi(\pi_k|g)),\pi_k\in\Pi,k=1,2,\cdots\} \quad (2)$$

其中，g 为代理的决策意图。

代理不同决策意图下的收益记为 $r_g(\pi)$，以图 1 的线性函数[3]定义 $A'(\pi|g)$ 中的隶属度函数：

$$\mu_\Pi(\pi|g)=\frac{r_g(\pi)-v_1}{v_2-v_1} \quad (3)$$

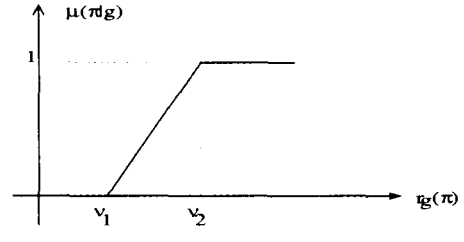

图 1　线性隶属度函数

式(3)表示代理在决策意图 g 选择策略组合 π 的可能性。

张化祥　博士，副教授，主要从事人工智能、机器学习、数据挖掘和分类器设计。黄上腾　教授，博导，主要从事 CIMS，数据库及数据挖掘技术研究。

$r_g(\pi)$表示决策意图为 g 的收益。定义代理的意图收益为在该意图下想要达到的收益，记为 $r(g)$。我们将代理在决策意图 g、执行策略组合 π 时的模糊收益记为 $r(g)\cdot\mu_\Pi(\pi|g)$。

定义 1(代理模糊收益)　设在一定决策心态下，代理有决策意图集合 $G=\{g_i, i=1,2,\cdots,n\}$。不同决策意图以多代理策略组合 Π 为论域的模糊集合定义如下：

$$A'(\Pi|g_i)=\{(\pi_k,\mu_\Pi(\pi_k|g_i)),\pi_k\in\Pi,k=1,2,\cdots\}(i=1,2,\cdots,n)$$

针对意图 g_i，代理意图收益为 $r(g_i)$。定义代理在此心态下的模糊收益为：

$$R=\sum_{i=1}^{n}\mathrm{sgn}(g_i)\cdot r(g_i)\cdot\mu_\Pi(\pi|g_i) \tag{4}$$

其中 $\mathrm{sgn}(g_i)$是以 g_i 为变量的符号函数。

代理决策心态决定决策意图。下面讨论不同心态下，代理策略组合隶属度及模糊收益的计算。自利心态下，代理决策意图只有一个，就是最大化自身收益。由式(3)计算代理的策略隶属度为：

$$\mu_\Pi(\pi|g)=\frac{r_g(\pi)-\min}{\max-\min} \tag{5}$$

其中 min 和 max 分别为所有策略组合中代理可获得的最小收益和最大收益。代理在意图 g 时的模糊收益为：

$$R_s=\max\{r(\pi),\pi\in\Pi\}\cdot\pi_\Pi(\pi|g) \tag{6}$$

$r(\pi)$为代理在策略组合 π 下的收益。

合作心态下，代理意图是最大化系统中所有代理收益之和，而不仅仅是最大化自身收益。设系统中代理数为 n，代理具有 m 个决策目标，分别记为 $g_1,g_2,\cdots,g_m$，对应的代理收益分别记为 $r_{g1}(\pi),r_{g2}(\pi),\cdots,r_{gm}(\pi)$。针对不同目标的策略组合隶属度函数计算如下：

$$\mu_\Pi(\pi|g_i)=\frac{r_g(\pi)-\sum\min}{\sum\max-\sum\min} \tag{7}$$

其中$\sum$min 为系统中所有代理的最小收益之和，$\sum$max 为系统中所有代理的最大收益之和。合作心态下，决策目标 g_1，$g_2,\cdots,g_m$ 下代理的模糊收益为：

$$R_c=\sum_{i=1}^{m}[\max\{\sum_{i=1}^{n}r_i(\pi),\pi\in\Pi\}\cdot\mu_\Pi(\pi|g_i)] \tag{8}$$

其中 $r_i(\pi)$为代理 i 执行策略组合 π 时的收益。此时，符号函数皆为 1。

另外，还可讨论其它心态下代理的决策。如自利＋敌对，代理最大化自身收益，同时最小化对手收益；合作＋合伙，代理最大化自身收益，同时最大化对手收益；服务心态，代理最大化其它所有代理收益，而不关心自身收益；利它主义，代理最小化自身收益，最大化其它代理收益。

3　模糊收益下代理策略的计算

给定策略组合 π，代理可以计算出其收益。如针对图 2 所示的两代理重复矩阵对策。

代理 1$\begin{pmatrix} r_{11} & r_{12} \\ r_{21} & r_{22} \end{pmatrix}$　代理 2$\begin{pmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{pmatrix}$

图 2　两代矩阵对策

我们记　$r_1=\max\{r_{ij}, i=1,2;j=1,2\}$；

$r_2=\min\{r_{ij}, i=1,2;j=1,2\}$

$c_1=\max\{c_{ij}, i=1,2;j=1,2\}$；

$c_2=\min\{c_{ij}, i=1,2;j=1,2\}$

假设两代理选择行动 1 的概率分别为 α 和 β，则选择行动 2 的概率分别为 $1-\alpha$ 和 $1-\beta$。此时，两代理的收益计算如下：

$$f_1(\alpha,\beta)=r_{11}\alpha\beta+r_{12}\alpha(1-\beta)+r_{21}(1-\alpha)\beta+r_{22}(1-\alpha)(1-\beta) \tag{9}$$

$$f_2(\alpha,\beta)=c_{11}\alpha\beta+c_{12}\alpha(1-\beta)+c_{21}(1-\alpha)\beta+c_{22}(1-\alpha)(1-\beta) \tag{10}$$

自利心态下，我们有 $r_g(\pi)=f_1$。隶属度函数计算为：

$$\mu_\Pi(\pi|g)=\frac{f_1-r_2}{r_1-r_2} \tag{11}$$

合作心态下，有 $r_{g1}(\pi)=f_1$ 和 $r_{g2}(\pi)=f_2$，由式(7)分别计算出 $\mu_\Pi(\pi|g_1)$和 $\mu_\Pi(\pi|g_2)$

$$\mu_\Pi(\pi|g_1)=\frac{f_1-r_2-c_2}{r_1+c_1-r_2-c_2}$$

$$\mu_\Pi(\pi|g_2)=\frac{f_2-r_2-c_2}{r_1+c_1-r_2-c_2} \tag{12}$$

将式(11)和(12)的结果分别代入式(6)(8)，可以得到自利及合作心态下代理的模糊收益函数。如针对代理 1，我们有：

$$R_s=\frac{r_1}{r_2-r_2}(f_1-r_2) \tag{13}$$

$$R_c=\frac{r_1+c_1}{r_1+c_1-r_2-c_2}(f_1+f_2-r_2-c_2) \tag{14}$$

下面我们计算自利与合作心态下代理的模糊收益及行动策略。

自利情况下，可通过最大化代理模糊收益计算其行动策略。由式(13)得两代理模糊收益为：

$$R_s^{(1)}=\frac{r_1}{r_1-r_2}(f_1-r_2)\text{ 和 }R_s^{(2)}=\frac{c_1}{c_1-c_2}(f_2-c_2) \tag{15}$$

我们将 $R_s^{(1)}$ 对 α 求偏导数和 $R_s^{(2)}$ 对 β 求偏导数。使导数为 0 可计算出 α 和 β。得：

$$(\alpha^T,\beta^T)=(\frac{c_{22}-c_{21}}{\mu'},\frac{r_{22}-r_{12}}{\mu'})$$

其中，$\mu'=r_{11}-r_{12}-r_{21}+r_{22}$，$\mu'=c_{11}-c_{12}-c_{21}+c_{22}$

所得结果与传统对策基于代理个体理性假设所得的结果相同，即代理寻求 Nash 均衡点。我们知道，$\alpha,\beta\in[0,1]$。如果(α^T,β^T)落在二维单位为 1 的单元内，说明对策存在混合策略的 Nash 均衡点，否则，该均衡点将在单元的某个角上出现。

合作心态下，两代理的模糊收益可以由式(14)计算

$$R_c^{(1)}=R_c^{(2)}=\frac{r_1+c_1}{r_1+c_1-r_2-c_2}(f_1+f_2-r_2-c_2)$$

代理最大化各自的模糊收益，我们有：

$$\begin{cases}\max\dfrac{r_1+c_1}{r_1+c_1-r_2-c_2}(f_1+f_2-r_2-c_2)\\ 0\leqslant\alpha\leqslant1,0\leqslant\beta\leqslant1\end{cases}$$

上式是一个以 α、β 为变量的非线性规划。将 f_1 和 f_2 代入上式，整理后有：

$$\begin{cases}\max(\mu+\mu')\alpha\beta+d\alpha+d'\beta\\ 0\leqslant\alpha\leqslant1,0\leqslant\beta\leqslant1\end{cases} \tag{16}$$

其中　$d=r_{12}+c_{12}-r_{22}-c_{22}$，$d'=r_{21}+c_{21}-r_{22}-c_{22}$

针对式(16)的规划问题，当 $\mu+\mu'$、d 及 d'同时为正时，$(\alpha^T,\beta^T)=(1,1)$；同时为负时，$(\alpha^T,\beta^T)=(0,0)$；同时为 0 时，(α^T,β^T)为有效单元内或边界上的任意点。

4　代理策略的学习

针对两代理、两行动的一般和重复对策，Singh 等[5]提出策略学习算法 IGA，并证明该算法能够保证策略收敛到 Nash 均衡点或代理迭代步上收益的平均值收敛到 Nash 均衡点处代理的期望收益。但代理迭代收益序列不一定收敛到 Nash 均衡点处代理的期望收益。为解决 IGA 算法存在的问题，Bowling[4]提出可变学习率的策略学习算法 WOLF-IGA，但

该算法并不能解决所有情况下的策略学习收敛[4]。为此，Banerjee 等[6]提出一种新的判断准则，解决了 WOLF-IGA 存在的问题。我们提出并利用基于冲量的学习算法 M-IGA[8]学习代理策略。实验表明[8]，M-IGA 保证代理策略收敛到 Nash 均衡点。针对前面的模糊收益，我们只讨论自利心态下代理策略的学习。

针对图 2 所示的矩阵对策，由式(15)知，自利心态下两代理的模糊收益由函数 f_1 和 f_2 完全决定。

由式(9)、(10)，我们得：

$$\frac{\partial f_1}{\partial \alpha}=\mu\beta-(r_{22}-r_{12}) \tag{17}$$

$$\frac{\partial f_2}{\partial \beta}=\mu'\alpha-(c_{22}-c_{21}) \tag{18}$$

代理沿着其期望收益增加的方向更新策略。我们有：

$$\begin{cases}\alpha_{t+1}=\alpha_t+\eta\dfrac{\partial f_1(\alpha_t,\beta_t)}{\partial \alpha}\\ \beta_{t+1}=\beta_t+\eta\dfrac{\partial f_2(\alpha_t,\beta_t)}{\partial \beta}\end{cases} \tag{19}$$

η 为学习的步长，当 $\eta\xrightarrow{t\to\infty}0$ 时，该算法即为 IGA 算法。

Bowling 改进了式(19)的更新规则。其思想是当代理“赢”得收益时，策略更新得慢一些，“输”掉收益时，策略更新快一些，但仍存在策略收敛问题。

策略学习与神经网络中权重学习[7]的实质相同，即在值函数增加的方向上增加策略。在值函数增加的过程中，策略更新的每一步都具有一定的试探性。如果在连续两步的迭代中都能保持值函数的增加，且在其中的后一步中值函数增加的速度快于前一步时，我们是否可以考虑增加学习的步长；其它情况，我们保持步长不变。基于以上思想，我们提出通过在策略迭代中增加冲量的方法学习代理的策略，并称之为冲量学习算法 M-IGA。此时，式(19)的学习规则变为：

$$\begin{cases}\alpha_{t+1}=\alpha_t+\eta\dfrac{\partial f_1(\alpha_t,\beta_t)}{\partial \alpha}+\delta_\alpha(t)\gamma\Delta\alpha_t\\ \beta_{t+1}=\beta_t+\eta\dfrac{\partial f_2(\alpha_t,\beta_t)}{\partial \beta}+\delta_\beta(t)\gamma\Delta\beta_t\end{cases} \tag{20}$$

其中 $\delta_\alpha(t)$、$\delta_\beta(t)\in\{0,1\}$，为一个大于零的常数，称为冲量。其中 $\Delta\alpha_t=\alpha_t-\alpha_{t-1}$ 及 $\Delta\beta_t=\beta_t-\beta_{t-1}$。由式(19)，我们可以很容易得到：

$$\begin{cases}\Delta\alpha_t=\eta\dfrac{\partial f_1(\alpha_{t-1},\beta_{t-1})}{\partial \alpha}\\ \Delta\beta_t=\eta\dfrac{\partial f_2(\alpha_{t-1},\beta_{t-1})}{\partial \beta}\end{cases} \tag{21}$$

针对 $\delta_\alpha(t)$、$\delta_\beta(t)$，我们提出如下取值规则：

$$\delta_\alpha(t)=\begin{cases}1 & \left.\dfrac{\partial^2 f_1(\alpha_t,\beta_t)}{\partial t^2}\right|_\beta>0\\ 0 & \text{otherwise}\end{cases}$$

$$\delta_\beta(t)=\begin{cases}1 & \left.\dfrac{\partial^2 f_2(\alpha_t,\beta_t)}{\partial t^2}\right|_\alpha>0\\ 0 & \text{otherwise}\end{cases} \tag{22}$$

式(22)的规则可直观表示如下：

$$\left.\frac{\partial^2 f_1(\alpha_t,\beta)}{\partial t^2}\right|_\beta=\frac{\partial}{\partial t}\left(\frac{\partial f_1}{\partial \alpha}\cdot\frac{\partial \alpha}{\partial t}\right)\Big|_\beta=\left(\frac{\partial^2 f_1}{\partial^2 \alpha}\cdot\left(\frac{\partial \alpha}{\partial t}\right)^2\right)\Big|_\beta+\left(\frac{\partial f_1}{\partial \alpha}\cdot\frac{\partial^2 \alpha}{\partial t^2}\right)\Big|_\beta$$

由式(19)，我们知道 $\dfrac{\partial^2 f_1}{\partial^2 \alpha}=0$，所以上式变为

$$\left.\frac{\partial^2 f_1(\alpha_t,\beta_t)}{\partial t^2}\right|_\beta=\left.\frac{\partial f_1}{\partial \alpha}\cdot\frac{\partial^2 \alpha}{\partial t^2}\right|_\beta$$

将 f_1 函数用式(17)代入，整理后可得：

$$[(\alpha_t-\alpha_{t-1})-(\alpha_{t-1}-\alpha_{t-2})][\mu\beta+(r_{12}-r_{22})]>0$$

由式(17)、(19)得到 $\mu\beta+(r_{12}-r_{22})=\dfrac{\partial f_1}{\partial \alpha}=\Delta\alpha_t/\eta$。其中 η 为一个大于 0 的无限小数，所以说 $\mu\beta+(r_{12}-r_{22})$ 与 $\Delta\alpha_t$ 同号。有：由 $[(\alpha_t-\alpha_{t-1})-(\alpha_{t-1}-\alpha_{t-2})][\mu\beta+(r_{12}-r_{22})]>0$ 可得 $[(\alpha_t-\alpha_{t-1})-(\alpha_{t-1}-\alpha_{t-2})]\Delta\alpha_t>0$

同理，对于代理 2，$\left.\dfrac{\partial^2 f_2(\alpha_t,\beta_t)}{\partial t^2}\right|_\alpha>0$ 的实质为：

$[(\beta_t-\beta_{t-1})-(\beta_{t-1}-\beta_{t-2})][\mu'\alpha+(c_{21}-c_{22})]>0$。同样我们有 $\mu'\alpha+(c_{21}-c_{22})$ 与 $\Delta\beta_t$ 同号。

文[8]的式(8)(9)构成了策略学习算法 M-IGA 的核心。文[8]的定理 2 给出了 M-IGA 学习算法收敛的证明。

小结 我们提出了多代理模糊收益的概念。传统对策论假定代理对每个行动策略具有相同的偏好，导致对策中代理学习 Nash 均衡点。实际上，代理对各行动策略存有一定偏好，且该偏好不能被准确描述。我们借助模糊集合理论，建立了联合行动策略域上的模糊集合，定义不同决策心态和意图下的策略隶属度及相应的模糊收益。

收益函数是代理学习其行动策略手段，我们利用 M-IGA 算法学习代理的策略。算法通过增加冲量项的办法保证代理策略的收敛。

BDI 模型给出了代理的信念、期望及意图逻辑框架。在定义模糊集合时，我们为简化问题，只考虑意图对隶属度函数的影响。实际上，代理信念、期望都影响着隶属度及模糊收益的定义。同时策略学习中，只考虑了代理自身收益变化对策略的影响，没考虑其它代理收益变化对策略学习的影响。这些都值得进一步研究。

参 考 文 献

1 Rao A S, Georgeff M P. Modeling rational agents with a BDI-architecture[A]. In: Proc. of 2nd Intl. conf. on principles of knowledge representation and reasoning, San Mateo CA, Morgan Kaufmann, 1991. 473～484

2 张平安，高春华，等译. 神经—模糊和软计算. 西安交通大学出版社，2000. 8～63

3 Song Q, Kandel A. A fuzzy approach to strategic games. IEEE Tran. on fuzzy systems, 1999, 7(6): 634～642

4 Bowling M, Veloso M. Multiagent learning using a variable learning rate. Artificial Intelligence, 2002, 136: 215～250

5 Kearns S M, Mansouv Y. Nash covergence of gradient dynamics in general-sum games. In: proc. of the 17th conf. on uncertainty in artificial intelligence, 2000. 541～548

6 Banerjee B, Peng J. Convergent gradient ascent in general-sum games. In: Proc. of the 13th European Conf. on Machine Learning, Printed by LNCS, 2002. 1～9

7 Mictchill T M. Machine Learning. The McGraw-Hill Companies, Inc. 1997

8 Zhang Huaxiang, Huang Shangteng. Convergent Gradient Ascent with Momentum in General-Sum Games. Neurocomputing, 2004, 61: 449～454

9 Littman M L. Friend-or-foe Q-learning in general-sum games. In: 18th ICML, Williams college, MA, 2001. 332～328

10 Hu J, Wellman M P. Nash Q-Learning for General-Sum Stochastic Games. Journal of Machine Learning research, 2003, 1: 1～30

11 张化祥，黄上腾. 多代理最优响应 Q 学习及收敛性证明. 计算机科学，2004, 31(4): 96～98