2009－MPS－73 (15)
2009／3／5

拡張 BDI logic を用いた BDI と強化学習の統合について

高田司郎[†1] 新出尚之[†2]

従来の BDI logic は，Bratman の「意図の理論」を基にした既存 BDI モデルの範囲で，信念・願望・意図などの心的状態とそれら心的状態を保持・更新して目的を達成する振る舞いに関して，形式的な議論や証明を行うことができた．しかし動的な環境下の合理的エージェントの実現には，強化学習との統合などが要請される．そこで，確率的遷移と不動点オペレータの概念を導入して BDI logic を拡張した $\mathcal{TGMSATG}$ を用いて，強化学習で用いられる方策や有限 MDP を $\mathcal{TGMSATG}$ の論理式として記述し，BDI と同じ論理体系で扱うことを可能にすることで，BDI と強化学習の統合方式を提案する．具体的には，強化学習の事例として「カヌーレーシング」を $\mathcal{TGMSATG}$ を用いて形式的に記述することで，厳密な議論や証明ができることを例示し，上記のように拡張された合理的エージェントの実現に，$\mathcal{TGMSATG}$ が有効であることを示す．

Integration of BDI and Reinforcement Learning Using An Extended BDI logic

SHIRO TAKATA[†1] and NAOYUKI NIDE[†2]

Using traditional BDI logics, within the existing BDI model which based on the theory of intention by Bratman, we can formally argue or prove various properties of agents' mental states such as beliefs, desires and intentions, or behaviors of agents to achieve their aims while holding and updating their mental states. However, to construct rational agents under dynamic environments, additional capabilities such as integration with reinforcement learning are required. In this paper, we describe the notions used in reinforcement learning, such as policies and finite MDPs, as a formula of $\mathcal{TGMSATG}$, an extended BDI logic with probabilistic transitions and fixpoint operators. In this way, we propose a way to integrate BDI and reinforcement learning by enabling us to handle those two within a uniform logic. Specifically, using $\mathcal{TGMSATG}$, we provide a formal description of canoe racing as a case of reinforcement learning, and give some examples of strict arguments and proofs. It shows the effectiveness of $\mathcal{TGMSATG}$ on realizing rational agents extended in the way described above.

1. はじめに

従来の BDI logic は，Bratman の意図の理論[1] を基にした既存 BDI モデルの範囲で，信念・願望・意図などの心的状態とそれら心的状態を保持・更新して目的を達成する振る舞いに関して，形式的な議論や証明を行うことができた．しかし動的な環境下の合理的エージェントの実現には，強化学習との統合などが要請される[3]．そこで，新出は確率的遷移と不動点オペレータの概念を導入して BDI logic を拡張した演繹体系 $\mathcal{TGMSATG}$ を提案した[4]．

本稿では，BDI logic を拡張した $\mathcal{TGMSATG}$ を用いて，有限マルコフ決定過程などの既知モデル上の強化学習と BDI エージェントの統合を再度試みる．具体的には，我々が従来から強化学習との統合のテストベッドとしている「カヌーレーシング問題」を使用して，従来の BDI logic では記述できなかった確率的遷移を記述できるように拡張した $\mathcal{TGMSATG}$ を用いて，BDI と強化学習とを統合した記述を例示することで，BDI logic を拡張した $\mathcal{TGMSATG}$ の有効性を示す．

†1 近畿大学理工学部
School of Science and Engineering, Kinki University
†2 奈良女子大学理学部
Facluty of Science, Nara Women's University

2. カヌーレーシング問題の有限 MDP

図 1 左のグリッドは川，上から下方向が川の流れ，△ は岩を表す．各マスを状態の単位とし，左から i 列目，上から j 段目のマスの状態を s_{ij} で表す．スタート地点は初期状態 s_{31} とする．ゴール地点は最下段の全てのマスであり，今回の目標はそれらのいずれかのマスに速くたどり着くことである．

次にカヌーレーシング問題の方策 $\pi(s_{ij}, a)$ について述べる．各状態 s_{ij} にいるときの行為 a はカヌー

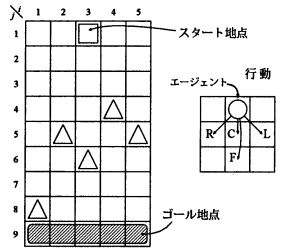

図 1 カヌーレーシングのグリッドワールド
Fig. 1 Gridworld of canoe racing

の漕ぎ方であり，行為の集合 $\mathcal{A}$ は $\{\mathrm{L}, \mathrm{C}, \mathrm{R}, \mathrm{F}\}$ とし，それぞれ左に移動，直進，右に移動，および，速く直進する．ただし，グリッドの左右へはみ出る方向へ移動する行為はとれず，また，最下段を通過するような行動をとった場合は最下段に留まるものとする．たとえば，川の流れが速い場合と遅い場合の環境モデルに関して，s_{34} における方策と有限 MDP を，それぞれ**表 1** と**表 2** で与える．

表 1 状態 s_{34} における方策
Table 1 Policy at state s_{34}

状態	行為	流れが速い場合の π	流れが遅い場合の π'
s_{34}	R	0.7	0.2
s_{34}	C	0.1	0.6
s_{34}	L	0.1	0.1
s_{34}	F	0.1	0.1

表 2 状態 s_{34} における有限 MDP
Table 2 Finite MDP at state s_{34}

	流れが速い場合			流れが遅い場合		
行為	確率	次状態	報酬	確率	次状態	報酬
R	0.5	s_{24}	-5	0.8	s_{24}	-1
R	0.5	s_{35}	-1	0.2	s_{35}	-1
C	0.9	s_{35}	-5	0.9	s_{35}	-1
C	0.1	s_{45}	-1	0.1	s_{45}	-1
L	1.0	s_{34}	-5	1.0	s_{34}	-5
F	1.0	s_{35}	-5	1.0	s_{35}	-5

次にカヌーレーシング問題の報酬について述べる．目標である最下段のマスが次状態であるときは，報酬 100 を得て，移動後，スタート地点 s_{31} へ戻される．また，△ のマスに移動，あるいは通過するような行動をとった場合は，元の状態，または △ のマスにぶつかる直前のマスに戻り報酬 -5 を得る．それ以外の移動では常に報酬 -1 を得るとする．たとえば，状態 s_{34} において，F の行動をとった場合は，流れが速い場合も遅い場合も確率 1.0 で s_{36} の岩に当たり，通過した s_{35} に留まり，報酬 -5 を得る．

3. 拡張 BDI logic $\mathcal{TOMAIO}$

本節では新出が提案した様相論理体系 $\mathcal{TOMAIO}$ (Theory about Observations of Multi-Agents with Tense and Odds) の構文と遷移確率に関する推論規則のみを紹介する．

3.1 論 理 式

3.1.1 構 文

まず $\mathcal{TOMAIO}$ の論理式の定義を与える．以下では，単に「論理式」と言えば $\mathcal{TOMAIO}$ の論理式を指すものとする．

一階言語と，イベント定数記号の集合 $\mathcal{E}$，エージェント定数記号の集合 $\mathcal{A}$ を各 1 つずつ選んで固定しておく．但し $\mathcal{E}$, $\mathcal{A}$ は有限集合とする．

- 一階述語論理の原始論理式は ($\mathcal{TOMAIO}$ の) 論理式
- ϕ, ψ が論理式ならば $\phi \vee \psi$, $\neg\phi$ も論理式 ($\wedge$, $\supset$ などは略記として導入)
- ϕ が論理式，x が変数記号ならば $\forall x\phi$ も論理式
- ϕ が論理式，$e \in \mathcal{E}$，$0 \leq p \leq 1$ ならば $\mathsf{X}^e_{\geq p}\phi$ も論理式 (従来の BDI logic の拡張)
- ϕ が論理式，$a \in \mathcal{A}$ ならば $\mathsf{BEL}^a\,\phi$, $\mathsf{DESIRE}^a\,\phi$, $\mathsf{INTEND}^a\,\phi$ も論理式

また，$\mathsf{AX}^e\,\phi$ は $\mathsf{X}^e_{\geq 1}\phi$ の略記とし，不動点オペレータについては割愛する．

3.1.2 直感的な解釈

直感的には，$\mathsf{X}^e_{\geq p}\phi$ は「イベント e を実行すると，1 時刻後に p 以上の確率で ϕ が成り立つ」を表し，CTL の next time オペレータ AX に相当するものの拡張である．なお，$\mathsf{X}^e_{\leq p}\phi$ は $\mathsf{X}^e_{\geq 1-p}\neg\phi$ の，$\mathsf{X}^e_{<p}\phi$ は $\neg\mathsf{X}^e_{\geq p}\phi$ の，$\mathsf{X}^e_{>p}\phi$ は $\neg\mathsf{X}^e_{\geq 1-p}\neg\phi$ の，また $\mathsf{X}^e_{=p}\phi$ は $\mathsf{X}^e_{\geq p}\phi \wedge \mathsf{X}^e_{\leq p}\phi$ の略記として書けるため，オペレータは $\mathsf{X}^e_{\geq p}$ だけ用意すれば十分である．

3.1.3 X オペレータに関する推論規則例

$$\frac{\phi, \psi \rightarrow \quad \psi, \xi \rightarrow \quad \xi, \phi \rightarrow}{\mathsf{X}^e_{\geq .5}\phi, \mathsf{X}^e_{\geq .5}\psi, \mathsf{X}^e_{\geq .5}\xi \rightarrow} \qquad \frac{\phi \rightarrow}{\mathsf{X}^e_{\geq .5}\phi, \mathsf{X}^e_{\geq .5}\psi, \mathsf{X}^e_{\geq .5}\xi \rightarrow}$$

$$\frac{\psi \rightarrow}{\mathsf{X}^e_{\geq .5}\phi, \mathsf{X}^e_{\geq .5}\psi, \mathsf{X}^e_{\geq .5}\xi \rightarrow} \qquad \frac{\xi \rightarrow}{\mathsf{X}^e_{\geq .5}\phi, \mathsf{X}^e_{\geq .5}\psi, \mathsf{X}^e_{\geq .5}\xi \rightarrow}$$

図 2 X オペレータに関する推論規則の例
Fig. 2 Example of inference rules about X operators

4. 有限 MDP の記述

この節では，従来の BDI logic と $\mathcal{TOMAIO}$ を用いてカヌーレーシング問題の有限 MDP を，それぞれ記述し比較する．

4.1 方　　策

表 1 において，状態 s_{34} の方策に関して，従来の BDI logic では確率が扱えないため，

$at(s_{34}) \supset \mathsf{A}(does(\mathrm{R}) \vee does(\mathrm{C}) \vee does(\mathrm{L}) \vee does(\mathrm{F}))$

のように，「いずれかの行為を実行する」ことしか記述できない．

一方 $\mathcal{TOMATO}$ を用いると，方策 $\pi_{s_{34}}$ は，以下のように，表 1 の方策そのものが記述できる．(ここで $true$ は，適当なトートロジーの略記とする)．

$$at(s_{34}) \supset \mathsf{X}^{\mathrm{R}}_{=0.7}\, true \wedge \mathsf{X}^{\mathrm{C}}_{=0.1}\, true \wedge \mathsf{X}^{\mathrm{L}}_{=0.1}\, true \wedge \mathsf{X}^{\mathrm{F}}_{=0.1}\, true$$

4.2 有限 MDP の記述

従来の BDI logic では，表 2 の有限 MDP の遷移確率と報酬に関して，

$$\mathsf{A}(at(s_{34}) \wedge does(\mathrm{R}) \supset \mathsf{X}((at(s_{24}) \wedge reward(-5)) \vee (at(s_{35}) \wedge reward(-1))))$$

のように，各行為に対する起こりうる遷移先の状態と報酬を記述することはできる．

一方 $\mathcal{TOMATO}$ を用いれば，表 2 の有限 MDP の遷移確率と報酬を下記のように厳密に記述することができる．

$$at(s_{34}) \supset \mathsf{X}^{\mathrm{R}}_{=0.5}(at(s_{24}) \wedge reward(-5)) \wedge \mathsf{X}^{\mathrm{R}}_{=0.5}(at(s_{35}) \wedge reward(-1)) \wedge \cdots$$

4.3 証明の例

行動 e_1 の実行後に $at(s_2)$ と $at(s_3)$ が排他的にしか成り立たないならば，2 つのいずれかが必ず成り立つことは，論理式 $\mathsf{X}^{e_1}_{\geq 0.7}\, p \wedge \mathsf{X}^{e_1}_{\geq 0.3}\, q \wedge \mathsf{AX}^{e_1}\, \neg(p \wedge q) \supset \mathsf{AX}^{e_1}(p \vee q)$ を証明することで示せる (ここで $at(s_2)$, $at(s_3)$ をそれぞれ p, q と略記した．また，$=$ でなく $\geq$ で十分である)．図 3 にその証明 (一部) を示す．

$$\dfrac{\dfrac{\dfrac{\vdots}{p, \neg(p \wedge q), \neg(p \vee q) \rightarrow}}{\dfrac{\vdots}{p, q, \neg(p \wedge q) \rightarrow}} \quad \dfrac{\vdots}{q, \neg(p \wedge q), \neg(p \vee q) \rightarrow}}{\dfrac{\mathsf{X}^{e_1}_{\geq 0.7}\, p, \mathsf{X}^{e_1}_{\geq 0.3}\, q, \mathsf{X}^{e_1}_{\geq 1}\, \neg(p \wedge q), \mathsf{X}^{e_1}_{>0}\, \neg(p \vee q) \rightarrow}{\dfrac{\vdots}{\rightarrow \mathsf{X}^{e_1}_{\geq 0.7}\, p \wedge \mathsf{X}^{e_1}_{\geq 0.3}\, q \wedge \mathsf{X}^{e_1}_{\geq 1}\, \neg(p \wedge q) \supset \mathsf{X}^{e_1}_{\geq 1}(p \vee q)}}}$$

図 3 証明の例 1
Fig. 3 Example of proof (1)

5. BDI と強化学習の統合

5.1 状況に適応した方策の選択

たとえば，表 1 では，状態 s_{34} においては，「流れが速い」環境モデルの方策 π と「流れが遅い」環境モデルの方策 π' を学習している．

今，BDI エージェントは流れの遅速を知覚することができるとし，スタート地点 s_{31} からゴール地点に向かうプランが，たまたま

(1) まず川の右岸 (図の最左行) に到達し，それから最下段へ向かう

(2) まず s_{34} に到達するか通過し，それから最下段へ向かう

の 2 つであったとする．

BDI モデルの枠組で「右岸に到達」のような副目標を持ったプランをどのように学習するかは今後の課題として，本稿では，たとえば下記のようなプランを人手で記述するものとする．

プラン 1:

- 意図形成条件: DESIRE(最下段に到達)
- 実行前提条件: スタートの合図
- 本体: 右岸に到達, 最下段に到達
- add list: 最下段に到達

プラン 2:

- 意図形成条件: DESIRE(最下段に到達)
- 実行前提条件: スタートの合図
- 本体: 状態 s_{34} に到達, 最下段に到達
- add list: 最下段に到達

「最下段に到達」という意図が形成された後，実際にスタートの合図があったとき，実行に移せるプランは，プラン 1，プラン 2 の 2 つである．いま，BDI エージェントがプラン選択のための基準として「流れが遅い場合は近道であるプラン 2 が有利だが，流れが速い場合は岩にぶつかりにくく流れに乗りやすいプラン 1 が有利」との信念を持っていたとする．

実際に流れが速いと，プラン 1 の方が選択され，本体の「右岸に到達」が意図として形成される．

5.1.1 従来の BDI logic を用いた方策の選択

BDI logic では，プラン 1 の選択状況は

$$\mathsf{DESIRE}(\mathsf{AF}\text{最下段到達}) \wedge \mathsf{BEL}(\text{流れ (速い)}) \supset \mathsf{INTEND}(\text{プラン 1})$$

という論理式で，また「右岸に到達」というサブプランを実行中の状況は

$$\mathsf{INTEND}(\mathsf{AF}\,\text{右岸到達}) \supset \mathsf{A}(\mathsf{INTEND}(\mathsf{AF}\,\text{右岸到達})\ \mathsf{U}\ (\mathsf{BEL}(\text{右岸到達}) \vee \neg\,\mathsf{BEL}(\neg\,\mathsf{EF}\,\text{右岸到達}))) \quad (1)$$

$$\mathsf{INTEND}(\mathsf{AF}\,\text{右岸到達}) \supset call\text{-}RL(args) \quad (2)$$

などの論理式で表現される．

ここで (1) は single-minded なコミットメント戦略[2] と呼ばれる，意図の持続性についての一般的な条件の 1 つである．

また式 (2) は，「右岸に到達」の意図がある間は毎

回，強化学習を呼び出して基本的行為を選択することを表す．ここで，*call-RL*(*args*) は強化学習部分を呼び出して次の基本行為の選択を決める副作用を持つ原始論理式で，*args* には現在の状況 (状態，流れの速さなど，学習済みのどの方策を用いるかの決定に必要なパラメータ) を渡す．強化学習側は，渡されたパラメータにより，学習済みの方策のいずれを用いるかを決定し，決定した行動を BDI 側に返せばよい．

5.1.2 $\mathcal{TOMATO}$ を用いた方策の選択

強化学習の学習が収束している場合は，4 節で記述した環境モデル別の方策を，プラン中に記述することが可能であるため，何らかの方法で，$\mathcal{TOMATO}$ を用いて記述された方策を解析して，確率的遷移として次の行為を決定することができる．つまり従来の BDI logic と異なり，*call-RL* を呼ぶことなく論理の枠組内のみで次の行為を実行することができる．

では，学習が収束していない場合はどうするかは今後の課題であるが，たとえば，強化学習は BDI と並行してオンライン学習を行ない，適当な頻度で，強化学習にて学習した方策を入力として，$\mathcal{TOMATO}$ を用いた記述を得る関数を作成することで，BDI エージェントがその方策に関するプランを更新するという方法が考えられる．

5.2 動的なプランの再熟考

まず従来の BDI logic の場合を述べる．式 (1) の，つの右辺の until の実行中に，右岸に到達しないうちに岩にぶつかってしまい，これを知覚した BDI アーキテクチャが，もはや右岸に到達できる見込みがないと判断したとする．

このとき，BEL(EF 右岸到達) は成り立たなくなり，式 (1) による INTEND(AF 右岸到達) の持続は終了するため，式 (2) で *call-RL*(*args*) が呼ばれることもなくなり，強化学習のスキルを使った行動は一旦ストップする．BDI エージェントはここで再度熟考して，別のプランを選択し直して実行することが可能である．また，プランの実行中に流速が変化し，現在実行中のプランを捨てざるを得ないと判断した場合にも，BDI は同様にプランの実行を中止して，別なプランに切り替えることができる．

次に，$\mathcal{TOMATO}$ を用いた場合を述べる．$\mathcal{TOMATO}$ は従来の BDI logic を含むため同様に，式 (2) の，つの右辺の until の実行中に，もはや右岸に到達できると見込みがないことは判断できる．つまり従来の BDI logic と同様に，強化学習機構だけでは困難な，動的な状況変化への柔軟な追随ができ，BDI と強化学習の両者の利点を行動決定に生かすことができる．

6. 考　　察

本稿では，従来の BDI logic と $\mathcal{TOMATO}$ を比較した．今後検討を要すると考えられる課題も多く，本節ではそのいくつかについて述べる．

6.1 BDI エージェントの実装

我々は，BDI エージェントを実際に実装することも視野に入れている．本稿で提案した $\mathcal{TOMATO}$ を用いて記述した方策と有限 MDP を実装したエージェントにどのように採り入れれば，5.1.2 で述べた $\mathcal{TOMATO}$ を用いた方策の選択が有効になるかは今後の重要な課題である．

6.2 遷移確率の使用方法

X オペレータに関する推論規則を用いて証明すべき有益な命題にどのようなものがあるのかは今後の課題である．特に，コミットメント戦略に関すると思われる予測行動価値を判断して，現在行っている意図を破棄すべきか推論することは重要である．

7. ま と め

本稿ではまず，確率的遷移と不動点オペレータの概念を導入して BDI logic を拡張した $\mathcal{TOMATO}$ を用いて，方策や有限 MDP を $\mathcal{TOMATO}$ の論理式として記述し証明を例示した．次に，$\mathcal{TOMATO}$ を用いた BDI と強化学習の統合方式を提案し，反射的行為の実行が柔軟に出来ることを示した．具体的には，強化学習の事例として「カヌーレーシング」を $\mathcal{TOMATO}$ を用いて形式的に記述することで，厳密な議論や証明ができることを例示し，上記のように拡張された合理的エージェントの実現に，$\mathcal{TOMATO}$ が有効であることを示した．

参 考 文 献

1) Bratman, M.E.: *Intention, Plans, and Practical Reason*, Harvard University Press (1987).
2) Rao, A.S. and Georgeff, M.P.: Modeling Rational Agents within a BDI-Architecture, *Reading in Agents* (Huhns, M.N. and Singh, M.P., eds.), Morgan Kaufmann, San Francisco, pp. 317–328 (1997).
3) 高田司郎，山川　宏，宮崎和光，新出尚之，長行康男，酒井隆道：強化学習と BDI の統合について—カヌー・レーシングを例題とした統合手法の考察，第 18 回人工知能学会全国大会論文誌 (2004). 1F1-02.
4) 新出尚之：確率的遷移と不動点オペレータを持つ BDI logic について，エージェント合同シンポジウム (JAWS2008) 講演論文集 (2008).