3F2-4

MDP 集団におけるマルチユーザ学習エージェント

Multi User Learning Agent on the Distribution of MDPs

片上 大輔[*1] Daisuke KATAGAMI
新田 克己[*1] Katsumi NITTA
宮崎 和光[*2] Kazuteru MIYAZAKI

[*1] 東京工業大学大学院 Tokyo Institute of Technology
[*2] 大学評価・学位授与機構 National Institution for Academic Degrees and University Evaluation

Recently, the researches of social robots and agents that smoothly communicate to human are paid to attention. The conventional researches seem to have aimed at the achievement of the society in the meaning of the wide sense. In a word, it is an approach where the designer embedded social competence like human. In this paper, we adduce the current state and the approach in the construction of such social robots and agents from the viewpoint of the information engineering and robotics, and discuss the methodology with a new design of social intelligence to aim at symbiosis and diversity with human.

1. はじめに

近年ロボットやエージェントのインタラクションに関する研究が注目されている[山田 02]。最近では特に「ロボットの社会性」という言葉が使われるようになってきた。この分野において、[Fong03]は、このテーマに関する貴重なサーベイである。ここにも、さまざまな社会性を目指したロボット研究が紹介されているが、現状では基本的に1対1におけるインタラクションを対象とした研究が多い。方法論としても人間のような社会的能力を設計者が埋め込み的に設計するアプローチが多く、そこでは広く人間とインタラクトするような広義の意味での社会性の実現を目的としているように思われる。

一方、人工物（エージェントやロボット）における知性の設計は、従来 AI の分野で行われてきた記号処理に基づく人工知能では限界があると言われている。近年のさまざまな研究により、身体性や、環境との相互作用に基づく知性の構築の重要性が認識されてきた。このような考え方は創発実体主義[國吉 03]とも呼ばれ、例えば認知発達ロボティクス[浅田 99]、身体性認知科学[Pfeifer01]などの学問分野がそれらに対応する。これらと脳科学の分野での知能探求の学問を総称してインテリジェンスダイナミクス（動的知能学）[土井 05]とも呼ばれている。人工物の真の人間らしさを目指すには、これらのアプローチのようにボトムアップ的に迫るような構成論的アプローチによる知性の獲得が望まれる.

前述のように、自律的な人工物が真の社会的知性を獲得するには、機械学習（machine learning）に代表されるような学習機構に、周りの主体との相互作用から自己組織化的に自己の行動を学習するようなボトムアップな学習機能を組み入れる必要があると考える。このようなボトムアップに創出する狭義の意味での社会性獲得にこそ、人工物における普遍的な人間らしさやそこにおける個性が生まれるものと筆者らは考えている。

本稿では、直接的経験とは別にユーザ間の類似性に応じて影響を受けるといった代理経験的な学習を行うエージェントである、モデルベースのマルチユーザ学習エージェント（M-MULA）を提案し、計算機上におけるマルチタスクシミュレーションによりそのインタラクションの効果を検証する。

連絡先：片上大輔，東京工業大学大学院総合理工学研究科，
〒226-8503 神奈川県横浜市緑区長津田町 4259 J2-53，
TEL&FAX: 045-924-5218，katagami@ntt.dis.titech.ac.jp

2. 社会的知能の実現

2.1 アプローチ

社会的知能実現のアプローチとして，我々は大きく2つのアプローチに大別されると考える．一つは設計者が社会的機能を埋め込み的に設計するトップダウンアプローチであり，もう一つは主体が環境や相手との相互作用から影響を受けることにより社会性を獲得するボトムアップアプローチである．例えば，環境の設計では，一般にトップダウン的に設計がされることが多いが，近年では，ロボットが持つ要素の設計またはインタラクション設計に関して，認知発達ロボティクス，インテリジェントダイナミクスなど，ボトムアップ的なアプローチが重要視されている．我々もボトムアップ的学習アプローチにより社会的集団における知を構築することを目標とする。

2.2 社会的環境と学習機能

学習という視点で、社会的環境を想定すると、以下のような幾つかの問題点が浮き上がってくる。Isbell らはチャット環境上の学習エージェントに対する多人数からの強化学習問題[Isbell 01]において、社会的環境における幾つかの問題をまとめている。これらを我々が修正したものを以下 4 つにまとめてみた。

(1) 複数の報酬系の存在： 異なった多くのユーザが介在する異なった報酬系から報酬が与えられるため、系によっては矛盾していたりすることもある。これをどう統合し、どう学習させるのかが難しい問題となる。

(2) ユーザからの報酬の不一致性： ユーザには様々な特性（嗜好、意図など）があり、一般に相互作用を通して個別に獲得する必要がある。

(3) 各ユーザの報酬の変動性： 各ユーザからの報酬もインタラクションを長く続けているとユーザの特性（熱心さ、飽き易さ、嗜好、感情の変化など）により同じ状況においても報酬が変動してしまう。これが学習の収束を難しいものにしている。

(4) 訓練の希薄性： 現実には、各ユーザと十分な訓練を行うことは困難であり、十分な訓練を想定することは現実的に

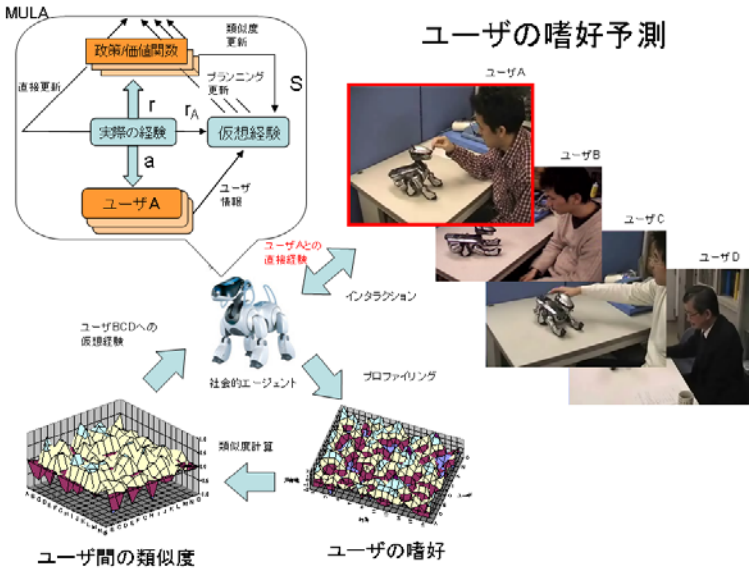

図 1 Multi User Learning Agent[片上 05]

有用なシステムとはいえない。このスパーシティ(sparsity)の問題を克服する必要がある。

社会的環境で学習を効率的に行なうには、これらの問題に対して対処することが必要である。

(1) 複数の報酬源の問題に関しては、ユーザの報酬系の分布を定量的に定義し、そこからの学習を考える必要がある。また、局所的に各ユーザからの報酬を最大化するとともに、長期的にみてエージェントの一生を通じての報酬を最大化することが目標となる。この問題は[田中 03]でも扱われているが今回は主に扱わないものとする。

(2) ユーザからの報酬の不一致性に関しては、各ユーザとのインタラクションを通して、個別にモデルを作成し各モデルを利用することで適応することが可能である。

(3) 各ユーザの報酬の変動性に関して、作成された各モデルへの適応を動的に学習することで適応することができる。

(4) 訓練の希薄性に関して、モデルベースアプローチを採用することで、直接ユーザとのインタラクションと平行して、各環境に対応する各モデルから仮想的に最適政策のプランニングを行なう手法が知られている[Sutton 90]。これを採用することで直接経験と仮想的経験を利用し、オンラインで十分な経験を行なうことが期待できる。

我々はこれまでに、複数ユーザとのインタラクションの類似性から各ユーザとの学習を高速化する手法である社会的学習エージェント(MULA)を提案を行なってきた[片上 05]。図 1 に MULA を利用したユーザの嗜好予測実験の概要を示す. MULA はユーザ間の類似度から代理経験的に学習を行うことで, あまりインタラクションがないユーザにも効率的に学習できるエージェントである.

本稿では、この MULA をベースに、正式にモデルベースアプローチを採用し、上記の(1)を除く3つの問題に対して対処の可能性を探ることを目的とする。定式化されたマルコフ決定過程(MDP)の上のマルチタスク強化学習問題に適用し、定量的にその効果を検証することにする。これらの議論は対象問題を明確にするだけではなく、今後社会的環境における学習の問題に対して研究を行っていくためにも非常に重要であると我々は考える。

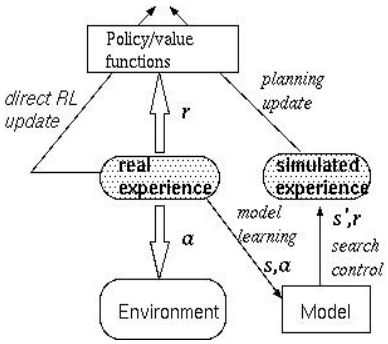

図 2 Dyna アーキテクチャ

3. マルチタスク学習問題

集団への適応問題の関連として, 強化学習の分野ではマルチタスク学習問題が関連深い. ここでの目標は, 全てのタスク内においてそれ以前のタスク学習で得られた経験を用いることにより一定期間内における総獲得報酬の最大化を目指すことである[田中 03]. このために, エージェントはなるべく少ない行動ステップ数でより良い次状態の価値の推定をする必要がある. 一般に強化学習では, 一回の行動の負荷が高い場合, モデルベースと呼ばれるアプローチを採用する場合が多い. これは通常のオンライン的強化学習(実際の経験)に加えて, 同時にそこで得られた経験からタスク毎に外界のモデルをエージェント内に構築し, そこにおける仮想的強化学習(仮想経験)を行うことで, 単一の価値関数推定を加速しようとする方法論(図 2)である. 最も基本的なモデルベース強化学習(Dyna-Q algorithm [Sutton 90])の具体的な手順を図 3 に示す. ここでは, 通常の Q-learning と, モデルの更新, そして X 回のモデル内 Q-learning(最適方策へのプランニング)を続けて行っている.

田中らは Dyna-Q に統計量(平均と標準偏差)と過去経験を利用したマルチタスク強化学習エージェントの実現を行っている[田中 03]. ここでは, Value パラメータを複数タスクとの経験から適応し構築していく試みを行っており複数タスクに適応した研

1. Initialize q(s,a,i) for all s,a,i
2. Do forever:
1) User identification $i \in I$
2) s←the current state.
3) a←policy(s,q).
4) r,s' ←carry out action a.
Direct learning
5) Simple Q-learning
Indirect Larning
6) Model(s,a)←s',r.
7) Repeat X times (do planning)
(a) s,a ← random selection from experienced states and actions
(b) r,s' ←Model(s,a)
(c) $q(s,a) \leftarrow q(s,a) + \alpha(r + \gamma \max_{a'} q(s',a') - q(s,a))$

図 3 Dyna アルゴリズム[Sutton90]

究であると言える. これらのマルチタスク学習エージェントにとって求められる能力とは、過去の経験を持ち続けて新たなタスク学習に活かすことである。ところが、マルチタスクを扱った RL 研究はまだ少なく、特に複数の人間とのインタラクションを指向した研究は少ない。

4. モデルベースドマルチユーザ学習エージェント

4.1 概要

ここでの目標は, ユーザとのインタラクションにおいてそれ以前のユーザとの学習で得られた経験を用いることにより対象のユーザへの適応を実現することである. 前述の社会的学習機能を備えたエージェントを本研究では, マルチユーザ学習エージェント(Multi User Learning Agent: MULA)と呼び, 本稿では、モデルベース強化学習(Dyna-Q アルゴリズム[Sutton 90])を基本として, 以下に説明する2つの学習方法によりモデルベースの M-MULA(Model-based Multi User Learning Agent)を実現する. マルチタスク学習では, 複数の環境に対応するために環境毎に知識やルールセットを構築しそれを再利用するアプローチが多く用いられる[Parr 98][田中 03]. 本稿では、それに習い各ユーザとのインタラクションに対してモデル構築し、それを他ユーザとの学習に利用する。

(1) **直接的学習とモデルの構築**(Direct Learning)

エージェントは、対象ユーザとの直接のインタラクションの経験より(1)式により直接更新を行なう。また同時に対象ユーザの

$$q(s,a) \leftarrow q(s,a) + \alpha(r + \gamma \max_{a'} q(s',a') - q(s,a)) \quad (1)$$

$$Model(s,a) \leftarrow s', r \quad (2)$$

状態sと行動 a、次状態 s'を用いることでそこで得られる報酬 r と状態遷移確率によりモデル構築を行なう((2)式)。

(2) モデルからの間接的学習(Indirect Learning)

ユーザからの報酬の不一致性に対応するためには, 各ユーザとの過去経験をうまく利用した対応が重要である. 対象ユーザとのインタラクション経験を利用し、モデルを作成することで、モデルからの間接的学習(プランニング)を行なうことができる. これにより、訓練を仮想的に行なうことができ、訓練の希薄性の問題にも対処することができる。構築されたユーザ毎のモデルから得られた次状態 s'と期待報酬 r((3)式)を用いて、(4)式により更新を行う.

$$r, s' \leftarrow Model(s,a,i) \quad (3)$$

$$q(s,a) \leftarrow q(s,a) + \alpha(r + \gamma \max_{a'} q(s',a') - q(s,a)) \quad (2)$$

(3) 過去経験(モデル)からの間接的学習(プランニング)
(Social Biased Indirect Learning)

複数のユーザに効率的に対応するためには, 複数のユーザとの過去経験をうまく利用した社会的な対応が重要である. 対象ユーザとのインタラクションは過去の類似ユーザとの経験が役に立つ. 特に対象ユーザとの経験が浅い段階では, インタラクションの指針になりうる. そこで, (2)式により構築されたユーザ毎のモデル間の類似性を用いて, 対象のユーザの学習を(5)および(4)式により行なう.

(4) 過去経験(モデル)の再利用(Social Biased Experience)

マルチタスク学習では、初期値の重要性が認識されている[田中 03]。ユーザが変わるたびに q 値を初期値から始めるので

$$r, s' \leftarrow Model(s,a,j) \quad (5)$$

$$q(s,a) \leftarrow q(s,a) + \alpha(r + \gamma \max_{a'} q(s',a') - q(s,a)) \quad (2)$$

はなくこれまでの経験を利用しすでにある程度効果が望めるような状態からの学習を行なうべきである。ここでは[田中 03]と同様、value 平均値((6)式)を用いる。

$$\overline{Q}(s,a) = \frac{1}{n}\sum_{i=1}^{n}(q(s,a)_i \quad (6)$$

4.2 M-MULA の学習手続き

M-MULA アルゴリズムを図 4 に示す. M-MULA ではまずユーザを認識し、そのユーザとのインタラクション(ある状態 s に対するユーザへの行為出力 a に対して報酬 r を受け取る)から simple Q-learning(1)式 を用いて, ユーザに個別の政策/価値関数を直接的に更新する(Direct learning)。同時に、対象ユーザの状態sと行動 a、次状態 s'を用いることでそこで得られる報酬 r と状態遷移確率 T により各ユーザに対するモデルを作成する((2)式)。

仮想的学習(Indirect Learning)では作成された現在のモデルを用いて仮想的に X 回のプランニングを行なう((3)式)。また、同様に作成された過去のモデルを用いて現在の対象ユーザのモデルとの類似度 を計算する. この類似度と現在のユーザとのインタラクションの情報 , s ,a を利用して, 間接的に過去のモデルから q 値を更新する((4)式).

この 2 つの経験により前節の直接的学習, 間接的学習を実現する. 本研究では, q 値の更新に社会的要素を導入しているといえる. ここで, q 値はユーザ j のモデルに対する状態 s、行為 a に対応し, α は係数, r はユーザ j のモデルからの報酬である。ここでは, ユーザ j のモデルとの経験により間接的にユーザ i の q 値を更新している.

5. 実験

5.1 実験目的

4 章の間接的学習を利用した社会的強化学習法を実装して、マルコフ決定過程(MDP)の上のマルチタスク強化学習問題に対する性能検証を行う。

1. Initialize q(s,a,i) for all s,a,i
   Employed mean value (6) or not.
2. Do forever:
1) User identification $i \in I$
2) s←the current state.
3) a←policy(s,q).
4) r,s' ←carry out action a.
  Direct learning
5) Simple Q-learning (1)
  Indirect Learning
6) Model(s,a,i)←s',r (2)
7) Repeat X times (do planning)
  (a) s,a ← random selection from experienced states and actions
  (b) r,s' ←Model(s,a,i) (3)
  (c) $q(s,a) \leftarrow q(s,a) + \alpha(r + \gamma \max_{a'} q(s',a') - q(s,a))$ (4)
8) Repeat X times (do planning from experienced models)
  (a) s,a←(random or roulette or all) selection from an experimented model j

図 4　M-MULA アルゴリズム

5.2 実験設定

実験のタスクとして図 5 に示すようなグリッドワールド上における迷路問題を用いた。これは、Sutton が Dyna システムの検証のために作成したものであり、ベンチマーク的問題となっている。始点(S)から終点(G)までの経路を学習する問題である。6x9 の各セルに7つの障害物があり、46+終点(G)の 47 状態が存在する。エージェントは常にその中のどれか1状態(セル)にあるものとする。エージェントは東西南北の 4 方向への移動が可能であり、決定的に状態遷移が行われる。エージェントは障害物や枠外へ遷移することはできない。終点状態(G)に達すると報酬値 100 が与えられて始点(S)へと戻される。あらかじめ終点(G)の位置を知らないエージェントは、終点で得られる報酬と試行錯誤により自律的に最短経路を獲得することを目的とする。

ここでは、ユーザをこの迷路(タスク)に見立てて、エージェントに対して複数の迷路問題が順番に与えられるという設定を考える。文献[田中 03]に倣い、上述の Sutton 迷路では決定的であった状態遷移を確率的なものに変更し、その確率値に変動を持たせることにより各タスク同士に違いを作成する。本実験ではユーザの報酬の不一致性を本迷路上に実装するために、簡易的に確率値をある正規分布から作成して、その正規分布の集合を形成する。エージェントに新たにタスクを与える際には、その集合を用いて全ての状態遷移確率値を独立に作成し、一つのタスクを作成する。状態遷移を確率的に行なうセル数は 5、25、40 とし、これらの場所はランダムで選択する。1実験で行なうタスク数は 10 とする。

5.3 比較対象と評価方法

学習は訓練例とテストフェイズを分けずにオンラインで追加的に行ない検証する。検証するための手法は以下の通り。

- Dyna-Q-Continue(Q-Table を複数環境を通して継続使用)
- M-MULA-random(経験済みのモデルからランダム選択し更新)

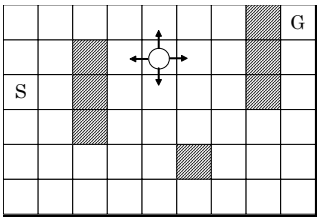

図 5　Maze problem (Sutton)

- M-MULA-roulette(経験済みのモデルから類似度に応じて roulette strategy により選択し更新)
- M-MULA-Best(経験済みのモデルのうち最大の類似度を持つモデルから更新)

評価方法は、全 q 値に関する(MDP 集団の各環境に対する最適 q 値との間の)平均自乗誤差(RMSE)を基本として評価を行なう。

6. おわりに

自律的な人工物が真の社会的知性を獲得するには, 機械学習(machine learning)に代表されるような学習機構に, 周りの主体との相互作用から自己組織化的に自己の行動を学習するようなボトムアップな社会的学習機能を組み入れる必要があると考える. 提案手法のようなボトムアップに創出する狭義の意味での社会性獲得にこそ, 人工物における普遍的な人間らしさやそこにおける個性が生まれるだろう.

参考文献

[山田 02] 山田誠二, 角所考:適応としての HAI, 人工知能学会誌, Vol.17, No.6, pp.658-664, 2002.

[Fong 03] T. Fong, I. Nourbakhsh and K. Dautenhahn: A survey of socially interactive robots, Robotics and Autonomous Systems, Vol.42, pp.143-166, 2003.

[國吉 03] 國吉康夫:ロボットの知能－創発実体主義の挑戦－, 計測自動制御学会誌計測と制御, Vol.42, No.6, 2003.

[浅田 99] 浅田稔, 石黒浩, 國吉康夫: 認知発達ロボティクスの目指すもの, 日本ロボット学会誌, Vol.17, No.1, 1999.

[Pfeifer 99] R. Pfeifer, C. Scheier: 知の創成－身体性認知科学への招待, 石黒章夫, 小林宏, 細田耕監訳, 共立出版.

[土井 05] 土井利忠:「ロボットは意識を持つか？」, 日本学術振興会学術月報, Vol.58, No.2, pp.92-95, 2005.

[Isbell 01] C. L. Isbell, C. R. Shelton, M. Kearns, S. Singh and P. Stone: A Social Reinforcement Learning Agent, Proc. of the Fifth International Conference on Autonomous Agent, 2001.

[田中 03] 田中文英, 山村雅幸:MDP 集団の上におけるマルチタスク強化学習, 電気学会論文誌 C, Vol.123, No.5, 2003.

[Sutton 90] R. S. Sutton, Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming, Proc. of the 7th International Conference on Machine Learning, pp.216-224, 1990.

[片上 05] 片上大輔, 大村英史, 安村禎明, 新田克己:社会的インタラクションに基づくマルチユーザ学習エージェント(MULA), 日本知能情報ファジィ学会誌, Vol.17, No.3, pp.340-350, 2005.

[Parr 98] R. Parr and S. Russel: Reinforcement Learning with Hierarchies of Machines, Advances in Neural Information Processing Systems 10, pp.1043-1049, 1998.