2A1-03

人工知能分野における強化学習研究の広がり

Current Researches of Reinforcement Learning in AI

山口 智浩
Tomohiro Yamaguchi

奈良工業高等専門学校 情報工学科
Department of Information Science, Nara National College of Technology

1. はじめに

これまでの人工知能分野での強化学習研究は，効用理論を基にした単一エージェント向けの学習手法，特に理論的解析が中心であった．しかしながら，1990年代での実問題への応用，特にマルチエージェントシステム（MAS と略す）への適用を通して，従来手法の枠組みの限界が明らかになりつつある．そこで本稿では，MAS 環境で他の学習エージェントとのインタラクションを扱うための知的エージェントの学習機能として，強化学習手法の枠組みをどう拡張すればよいかについて議論する．

2. 人工知能分野での強化学習研究の現状と課題

2.1 強化学習を研究する魅力

強化学習手法を研究する最大の魅力は，1）学習機能をもつ知的エージェント実現のための基本モデルとして最も単純であること，2）学習の適応的で柔軟な自律性について未解決の問題が残されていること，の2点である．

強化学習での学習目標および評価基準は，通常設計者である人間から与えられるため，これまでの大半の研究では，それらをエージェントが自身の行動する環境に対し，自律的にどう決定するか，という問題は議論されてこなかった．しかしながら，学習の視点を設計者からエージェントへ移して考えると，これらは強化学習効率を左右するだけでなく，環境の変動にどう能動的に適応するか，という重要な課題となる．さらに MAS は，学習エージェントにとって典型的な複雑，動的な環境である．そのため設計者が MAS 全体に与えた目標に対し，各エージェントがどう適切に各自の目標設定を行うかは，両者にとって困難な問題である．したがって，MAS における他エージェントとのインタラクティブな環境下で自律的エージェントを実現するためには，**目標，評価基準の自律生成**という問題を解決する必要がある．

2.2 最近 10 年間での主な成果

まず，最近の強化学習研究の流れを，学習手法，モデルの複雑化，マルチエージェント強化学習の3つに分けて，それぞれについて概要と課題について述べる．

(1) *Q* 学習からモデルベース手法へ

強化学習手法は，*Q* 学習に代表されるモデルレス手法がまず注目された後，最適政策への収束性や学習効率の良さから，MDP モデルに基づく手法に研究の主流が移ってきた．強化学習の目的は，単位行動当たりでの期待獲得報酬の最大化で，これを最大化する行動を各状態で与える政策を**最適政策**と呼ぶ．期待獲得報酬は政策に依存するので，主体は環境と相互作用しながら政策を探索することになる．この環境との相互作用をどう解釈するかで，強化学習研究は客観観測型と主観経験型との2種に分類できる．

客観観測型学習の**観測**とは，主体を環境から切り離した上で環境をモデル化することで，その特徴は，強化学習を政策に依存しない観測による環境モデル同定と，モデル上での最適政策の網羅的探索とに分けて行う点である．環境のモデル化を行うには，環境のクラスを仮定する必要がある．効用理論での動的計画法（DP）では，マルコフ決定過程（MDP）モデルでの最適化手法が知られていることから，観測結果から MDP モデルを統計的に同定し，最適政策を DP 法で探索する ad-DP 法が有名である．客観観測型は，環境のクラスが MDP やその類似クラスに限定される反面，最適性を追求できる利点がある．

これに対し主観経験型学習の**経験**とは，行動や政策に依存した学習結果を指す．主観経験型には，バケツリレー，Profit Sharing，*Q* 学習，Sarsa などモデルレスの強化学習法が相当する．手法によるが，学習結果が得られた経験の順序や分布に影響を受けるため，通常は最適解に収束しない．例えば*Q* 学習法が有限回で最適政策に収束しないのは，強化に用いる経験が実行した政策に依存し，かつ異なる政策を混合して学習するからである．主観経験型の主張は，主体の行動による相互作用は環境を変化させるので，客観観測，環境のモデル化は困難，という点である．したがって最適性を追求しない手法が多く，以降に述べるより複雑な問題を扱うためにモデルレス手法が再び注目されている．

(2) 単純 MDP モデルから POMDP モデルへ

単純 MDP モデルでの最適政策の探索手法は既に確立されていることから，1990 年代後半には，MAS に代表される現実的な問題で生じる**不完全知覚問題**を扱うための部分観測 MDP（POMDP）モデルでの強化学習手法に研究が移ってきた．

(3) 単一エージェントからマルチエージェント強化学習へ

従来の強化学習が，環境中に唯一存在する学習主体と外界との相互作用のみを扱ってきたのに対し，複数の学習主体で問題解決を行うマルチエージェント強化学習研究が盛んになってきた．但し，これまで成功した問題設定は，いずれも共通のゴールを持つエージェント間の協調タスクで，各エージェントが独立に獲得する報酬の最大化が全エージェントの報酬の最大化につながる問題である[荒井 2001]．

2.3 マルチエージェント強化学習の問題点と課題

MDP ベースの強化学習手法の問題点は，マルチエージェント強化学習への適用が原理的に困難なことである．その理由は，マルチエージェント間の相互作用によって生じる環境の動的性

連絡先：山口 智浩，奈良工業高等専門学校 情報工学科
〒639-1080 奈良県 大和郡山市 矢田町 22
Tel: 0743-55-6140, yamaguch@info.nara-k.ac.jp

と客観観測型学習とが両立しない世界だからである. すなわちMAS では主体同士が相互作用するので学習主体が客観観測する限りMASの構成要素とはなれず, 一方 MASの一員となって行動し他の主体と相互作用すれば, それが MAS 環境に影響を与えるため, 観測が主観的経験となるからである. さらに学習主体同士では相互のモデル化の無限退行という問題も発生する. つまり MAS 強化学習では, 各エージェントの視点で他の学習エージェントを含む環境を静的に同定するのが困難となる.

では, 上述の問題点を以下の3つの課題[荒井 2001]に分けて述べる. まず, **不完全知覚問題**とは, 他のエージェントの内部状態, 政策や行動が直接観測できないため, 外界が変動して確率的に知覚されることである. 次に**同時学習問題**とは, 複数の学習主体による相互の政策の変更が, 自己の行動による状態遷移先を変動させることにより生じる問題である. 第3に**報酬分配問題**とは, 目標達成に貢献した複数のエージェントに報酬をどう分配すべきか, という問題で, その論点は, 合理的な分配方法は存在せず, 何が公正かという価値観の問題であることである. つまり, 客観的な合理性の定義が困難なので, なんらかの形での主体間での利益分配や利害の調整, 価値観の形成機能が必要である. 以上3つの課題に共通する解決の方向性は, エージェント間で何らかの情報共有を行うことである.

3. 強化信号のコミュニケーションによる強化学習

強化学習に対する2つの立場として, 道具としての強化学習と, 知的エージェントの学習機能とが考えられるが, **報酬をいかに設定すべきか**, という論点は共通である. 本稿は後者の立場で, まずエージェントが自律的に報酬設定を行う必要性について述べ, 次にそれを実現するために強化学習手法の基本的枠組みをどう拡張すればよいか[山口 2000]について議論する.

3.1 社会的ジレンマ最小化による報酬分配問題の解決

社会的ジレンマとは, 各個体での自己利益の和が集団での利益と一致しない問題である. 言い換えると, 各強化学習エージェントの最適化学習である個人の自己利益追求が, MAS 全体の損失となり得ることでもある. これは, 集団挙動において個人の行動の貢献度をどう適切に評価するか, という信頼度割り当て問題に原因がある. 強化学習での獲得報酬和を効用値とした場合, MAS においては, 各個体の効用最大化と集団での効用最大化とが矛盾しないように学習目標である報酬を設定する必要があるが, 社会的ジレンマを生じないような事前の報酬設定が一般には容易ではない. そこで, 他のエージェントらとの相互作用をどう調整するかを扱う, 能動的かつ相互作用的な学習機能が必要となる. これを解決する有力な方法は, 各エージェントの自己利益最大化の和が集団での利益最大化となるように, つまり社会的ジレンマが小さくなるように, エージェントレベルでの目標設定を動的に調整することで, エージェントごとに独立な部分問題に分割し, 個別に強化学習することである.

3.2 強化信号のコミュニケーションによる利害調整

前節では, 社会的ジレンマを解決するためのマルチエージェント間での学習の相互作用的な調整機能の必要性を明らかにした. 次節では, 強化学習における学習目標を, MAS 間でコミュニケート可能な強化信号へと拡張することについて議論する.

3.3 強化信号と報酬との起源と役割

強化学習とは, 強化信号という学習のフィードバック情報を用いる機械学習法の一種である. **強化**(reinforcement)とは, 元々は, 動物の行動のメカニズムを刺激に対する反応で理解しようとした 1910 年代の行動主義心理学において, 動物が, エサ等の特定の刺激(**強化刺激**と呼ぶ)に対して, 次第に特定の反応を繰り返すようになることを指す. 行動科学分野では, 学習者の行動を左右する**強化刺激とは環境中の何か？**が追求されてきたのに対し, これまでの強化学習研究では, OR 分野での効用理論の影響のため, 報酬とは何か, といった基本的問題に関する議論が少なかった. しかしながら, 報酬から強化信号へと発想を転換すると, 環境中に存在する信号の一部を学習を方向付ける強化信号として利用することが可能となる.

3.4 学習目標の自律的探索に向けての課題

本節では, 学習の方向付けの自律性に関する課題と, 解決のアイデアについて議論する. まず, 学習の方向付けの自律性について, 以下の矛盾する2つの課題がある.

1)学習の客観性, 合理性を保証するには, 学習目標はエージェントの外部かつ固定であるべき.

2)外的な報酬・目標は, 学習者の内発的動機付けを低下させる.

これらを解決するため, 強化信号に基づくマルチエージェント強化学習の枠組み[山口 2000]を提案した. そこでは, 各エージェントが自他の学習目標を強化信号として生成, 発信し, エージェント間で強化信号を通信しながら,

1)自己の内発的な目標設定,

2)他者からの外的な強化刺激の利用,

との両者のバランスを相互に調整することによって上述の矛盾の解消を図る.

次に, 学習目標と評価基準の自律生成の課題について議論する. まず, 大局的外部目標が存在する場合には, それと矛盾せず, 社会的ジレンマを減少させる制約が, 各エージェントの自律的目標生成の合理的な評価基準であると考えられる. 一方, 大局的外部目標が存在しない場合には, 各エージェントの個別に探索, 生成する学習目標の生成, 評価基準が MAS 内で共有されることが, 各エージェントの獲得報酬和を最大化し, かつ共有された評価基準が進化的に安定であることを示せばよい.

4. おわりに

アメリカで発展してきた強化学習法や効用理論は, 最適化という名の元に, ある意味, 一人勝ちを追究する手法である. しかしながら, 生態学の知見によると, 有限の生態系では一人勝ちは過渡現象で系が安定せず, 系の安定点は**貸者の共存**となる. また, イモムシ, キャベツ, 寄生バチの3者生態系では, イモムシに食べられる際にキャベツの出す微量の化学物質(SOS 信号)が寄生バチを誘導する. 3者生態系のシミュレーションからSOS 信号の存在が緩衝となり, 3者系の安定範囲を広げることが示唆[Suzuki, 2002]されている. エージェントによる報酬の動的設定の課題は, 今後, マルチエージェント, 有限, という条件下で強化学習手法が発展していく上で, 重要であるといえる.

参考文献

[荒井 2001] 荒井幸代: マルチエージェント強化学習-実用化に向けての課題・理論・書技術との融合-, 人工知能学会誌, Vol.16, No.4, pp.476-481, 2001.

[Suzuki, 2002] 鈴木泰博, 高林純示, 田中博: 抽象化学系の3者生態系の解析, SICE 第 25 回システム工学部会研究会, pp.103-106, 2002.

[山口 2000] 山口 智浩, 強化信号のコミュニケーションに基づくマルチエージェント強化学習, (知能と複雑系 研究会) 情処研報, Vol.2000, No.66, 2000-ICS-121, pp.91-98, 2000.