1E5-5

多数のエージェントを利用した行動モデルの学習

An Analysis of Behavior Model Learning by Multiple Simulations

市瀬 龍太郎[*1] 森山 甲一[*2] 沼尾 正行[*2]
Ryutaro Ichise Koichi Moriyama Masayuki Numao

[*1]国立情報学研究所 大阪大学 産業科学研究所[*2]
National Institute of Informatics ISIR, Osaka University

We propose a machine learning method for generating behavior model. The method utilizes multiagent simulation approach. We conducted experiments with real simulation environment. The experimental results show that the proposed method successfully learns behavior model which has ability to avoid serious failure.

1. はじめに

人間は，さまざまな環境をセンシングし，行動の決定を行う．それと同じようなことを実現するために，人間の行動履歴から行動を模倣するプログラムを作成する行動クローニング (BC) [Sammut 96] の研究が行われてきた．しかし，実環境において，人間の行動は，すべての選択肢をカバーしているわけではない．本論文では，シミュレーション環境を用いることで，実際の人間の行動には表われないような行動についても，妥当性を調べながら，行動モデルを自動で学習する手法について述べる．そのような手法として，従来までは，シミュレーションの結果に応じて，行動を随時変更し，学習する手法が用いられていた [市瀬 11]．しかし，シミュレーション環境が動的な場合においては，シミュレーション結果と同等の結果が常に得られるとは限らないため，環境の影響を考慮する必要がある．そのため，シミュレーション手法を変更することにより，環境適性が調べられている [市瀬 12]．その過程において，学習された行動規則を用いると，ある特殊な状況にある場合に，大きく失敗する場合があることが明らかになった．従来の遺伝的アルゴリズムに基づく手法 [市瀬 11] では，学習過程において，うまく行動したものだけが良い行動規則として選択されるが，学習過程において遭遇しなかった特殊な状況においては，学習がうまくできないことがあるためである．そこで，本研究では，単一のエージェントの学習と多数のエージェントの行動からの学習を組み合わせた学習手法を提案し，その問題の解決を試みる．

2. 対象とするシミュレーション環境

本研究では，対象とするシミュレーション環境として，Happy Academic Life 2006(HAL2006) [山川 06] というゲーム型キャリアデザイン学習教材を用いた．HAL2006 は，人工知能学会 20 周年記念事業として開発された教育用ボードゲームで，プレイを通して，研究者のキャリアデザインを学習できるようになっている．プレイヤは自分のコマを進めながら，さまざまなイベントを疑似体験し，研究業績を積み上げて最終的なゴールを目指す．ゴールには，教育者型，悠々自適型，学内政治型，学術社会型，業績量産型，組織研究型，業績卓越型の 7 つがある．学習者は，プレイ途中で，体験するイベントにどのような判断をするかによって，自分の置かれる状況が変化する．そのため，プレイヤはさまざまな場面において，自分のゴールを達成するための適切な判断をしなければ，ゴールになかなか到達できないことになる．

HAL2006 は, 当初，紙を使ったボードゲームとして開発された．それを研究プラットフォームとして再構築し，電子化を行ったものが D-HAL2006 [市瀬 08] である．D-HAL2006 では，複数の人間の学習者が計算機を使ったプレイで，学習できるのみならず，人間の思考と同様な行動ルールを記述することで，人間の代わりに，エージェントがプレイすることもできるようになっている．本研究では，このシミュレーション環境を用いて，なるべく早くゴールすることができるような行動モデルを学習することとする．なお，現在，D-HAL2006 は，さまざまな修正を施された後,「Digital-HAL ver. 1.0」として誰でも利用できる形式で公開[*1] されている．

3. シミュレーションによる戦略学習

これまでに，遺伝的アルゴリズム (GA) を用いた進化計算によって，上記のゲームにおける戦略を獲得する手法が開発されてきた [市瀬 11]．その手法では，戦略ルールはそれぞれ遺伝子の列で表現され，各個体はその組合せからなっている．個体間の交叉と突然変異により複数の新個体を生成してシミュレーションを行い，その結果に基づいて個体を選択することを繰り返すことで，より早くゴールできる個体（ルールの組み合わせ）を発見する．しかし，単一のエージェント（個体）が経験できる状況は限られるため，これまでに経験をしていない状況になった場合に，うまく行動を選択できない場合がある．そこで，本研究では，複数のエージェントの行動履歴からエージェントの行動をさらに学習する手法を提案する．

学習が終わった多数のエージェント群に対して，シミュレーション環境中で動作を行わせると，エージェントが行動した行動履歴が得られる．当然，進化の過程で上位であったエージェントはうまく行動することが期待されるが，下位のエージェントであっても，うまく行動できることがある．そこで，うまく行動できた行動履歴を用いて，行動クローニングの手法を用いてエージェントの行動規則をさらに合成することを提案する．ここで用いた行動クローニングの手法は， [市瀬 09] の手法を用いた．このようなアプローチを取ることで，単一のエージェントで学習した場合と比べて，単一のエージェントが遭遇して

連絡先: 市瀬 龍太郎，国立情報学研究所情報学プリンシプル研究系，〒 101-8430 東京都千代田区一ツ橋 2-1-2，Tel:03-4212-2000，E-mail:ichise@nii.ac.jp

*1 http://www.academiclife.jp/

いない状況でうまく行動できる行動規則を含むことが期待できる．

4. 実験

まず，5 回のゲームでゴールするまでのターン数の合計を適合度とした GA で，100 個体 500 世代による実験を行った．ここでは，ゴールの種類によって戦略が異なるため，ゴールの種類毎に行動モデルを作成し，学内政治型と悠々自適型について検証することとした[*2]．得られた最良個体 100 個の行動モデルを用いて 100 回ゲームを行い，ゴールまでの平均ターン数と標準偏差を求め，平均ターン数順に並べた所，学内政治型では，表 1，悠々自適型では，表 2 の結果が得られた．表中央の GA による学習と書かれたのが，この結果であり，括弧内はその標準偏差を意味する．

次に，ここで得られた全てのエージェントの行動ログから，５０ターン以内にゴールできた行動ログのみを抽出し，行動クローニング (BC) の手法 [市瀬 09] を用いてカード選択の行動規則を学習した．そして，個々のエージェントが持つ行動規則の中で，カード選択の行動規則だけを入れ替えて，100 個の個体を生成し，それを 100 回実行した．表 1, 2 の BC による学習と書かれた部分がその結果である．

表 1 の学内政治型の結果を見ると，BC による学習を用いると，ゴールできない場合を大きく減らすことができている．GA による学習では，27 位以下のエージェントでは，ゴールに到達できない場合があったが，BC による学習では，全ての場合でゴールに到達できており，適切な行動規則を学習していることが分かる．また，標準偏差を比較すると GA による学習と比べて大きく抑えられていることが分かる．これは，冒頭で述べたような，これまでに経験されていない特殊な状況に陥った場合に，大きく失敗する問題を提案手法で解決できることを示している．一方，表 2 の悠々自適型の結果を見ると，上位のエージェントでは，GA による学習の方が少し優位であるが，下位では BC の方が優位となり，標準偏差は BC による学習の方が安定していることが分かる．専門家のプレイログを使って，同様な手法により行動規則の生成を行うと，ゴール到達平均ターン数は 45 となる [市瀬 09]．そのため，ここで GA により学習されている行動は，ほぼ最適なものであると言えるであろう．従って，BC による学習を行なっても，エージェント全体の結果が安定する効果は一部に見られるものの，直接的な学習効果は薄くなると考えられる．一方，表 1 の学内政治型では，学習が途中の状況であるため，BC によって，過学習を避けて，ゴールに到達できる行動規則を学習することができると考えられる．

5. おわりに

本研究では，シミュレーションを取り入れた行動モデル学習手法について述べた．シミュレーションを用いる場合には，人間の行動履歴から行動モデルを学習する手法に比べて，人間が遭遇しなかった状況に対しても学習を行うことができるという利点がある．しかし，学習過程において遭遇しなかった特殊な状況においては，学習がうまくできないことがある．そこで，本研究では単一のエージェントの学習結果を多数組み合わせて学習する行動クローニングの手法を提案し，その問題の解決をおこなった．今後は，別のゴールについても実験を行い，さらに詳細な検討を行う予定である．

*2 組織研究型についても同様に検証を行ったが，全てのエージェントが取った行動が同一だったため，ここでの報告は除外する．

表 1: ゴールまでの平均ターン数と標準偏差 (学内政治型)

	GA による学習	BC による学習
1 位	65.76 (38.50)	70.66 (31.19)
10 位	69.46 (37.13)	73.87 (31.29)
20 位	74.40 (36.27)	75.52 (33.95)
30 位	–	76.42 (38.99)
40 位	–	77.83 (34.05)
50 位	–	78.87 (39.10)
60 位	–	79.25 (35.32)
70 位	–	80.78 (40.34)
80 位	–	82.15 (37.60)
90 位	–	83.54 (40.46)
100 位	–	89.07 (43.97)

表 2: ゴールまでの平均ターン数と標準偏差 (悠々自適型)

	GA による学習	BC による学習
1 位	39.02 (6.08)	41.78 (7.64)
10 位	40.46 (8.71)	43.11 (8.78)
20 位	41.40 (7.29)	43.52 (8.63)
30 位	41.86 (8.15)	43.87 (9.60)
40 位	42.20 (10.28)	44.27 (8.97)
50 位	42.82 (9.10)	44.49 (10.22)
60 位	43.58 (10.22)	44.96 (9.02)
70 位	44.29 (18.11)	45.26 (8.46)
80 位	45.18 (9.83)	45.71 (10.27)
90 位	46.13 (12.16)	46.04 (10.76)
100 位	51.14 (15.99)	48.67 (7.43)

謝辞

本研究の一部は，物質・デバイス領域共同研究拠点における共同研究の支援により行われたものである．

参考文献

[Sammut 96] Sammut, C.: Automatic construction of reactive control systems using symbolic machine learning, *Knowledge Engineering Review*, Vol. 11, pp. 27–42 (1996)

[市瀬 11] 市瀬 龍太郎, 森山 甲一, 沼尾 正行：シミュレーション環境を用いた適切な行動モデルの学習, 第 25 回人工知能学会全国大会, 1G1-5 (2011)

[市瀬 12] 市瀬 龍太郎, 森山 甲一, 沼尾 正行：行動モデル学習における動的環境の影響, 第 26 回人工知能学会全国大会, 4B1-R-2-2 (2012)

[山川 06] 山川 宏, 市瀬 龍太郎, 太田 正幸, 加藤 義清, 庄司 裕子, 松尾 豊：Happy Academic Life 2006:研究者の人生ゲーム – ゲーム型キャリアデザイン学習教材の開発 –, 人工知能学会誌, Vol. 21, No. 3, pp. 360–370 (2006)

[市瀬 08] 市瀬 龍太郎, 庄司 裕子, 山川 宏, 三浦 麻子：学習者モデリング技術を用いたゲーム型教育システムのための研究プラットフォームの構築, 第 22 回人工知能学会全国大会, 2P2-12 (2008)

[市瀬 09] 市瀬 龍太郎, 山川 宏：ゲーム型教材のプレイ履歴からの行動知識の学習, 第 23 回人工知能学会全国大会, CD-ROM (2009)