1G1-5

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

Behavior Model Learning through Simulation

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 agent simulation approach. We conducted experiments with real simulation environment. The experimental results show that the performance of the proposed method is almost the same as behavior of experts in some cases.

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，機械学習の分野において，行動クローニング [Sammut 96] という方法を用いて，行動決定プログラムを自動的に作成する手法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ため，行動クローニングで学習されたプログラムでは，学習時にデータが得られなかった環境などにおいて，効率の悪い行動が学習されたり，想定外の行動が学習されることがある．本研究では，それを防ぐために，様々な行動をシミュレーションすることによって，人間の行動と同じように効率的に行動できるプログラムの作成を試みる．そのために，本論文では，進化計算を用いて，シミュレーション環境から，戦略を学習する手法を提案し，その結果の報告を行う．

図 1: D-HAL2006

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，目指す研究者に応じて，教育者型，悠々自適型，学内政治型，学術社会型，業績量産型，組織研究者型，業績卓越型がある．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．例えば，研究資金への応募を行うと，申請のために，論文を書いたりする研究時間を削らなければならないが，研究資金をあまり取らないと，資金不足で新たなテーマの研究ができなくなってしまうことがある．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(図 1) [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．

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

3. 進化計算による戦略の獲得

本研究では，遺伝的アルゴリズム (GA) を用いた進化計算によって，D-HAL2006 における戦略を獲得することを試みた．

3.1 プレイヤーエージェント

D-HAL2006 におけるプレイヤーの役割をするエージェントは，以下の 6 つのルール集合により，現在のコマ位置や所持時間などの状態を元に意思決定を行う．ただし，ルールに指定が無い場合にはランダムに意思決定を行う．

ゴール 最初のゴールおよびゴール変更イベントが生じた際のゴールの選択方法を決定

大学 10 の大学の選好順序を決定

学生 50 名の学生の選好順序を決定

ポスドク ポスドクを受け入れるかどうかと，20 名のポスドクの選好順序を決定

テーマ どの研究テーマを選択するかを決定

イベント ゲーム中で発生する下記の各種状況について，意思決定を規定

- ワイルドのマスの時にどのカードを選ぶか
- 適用が任意のカードを適用するか否か
- 教授公募に応募するか否か
- 転任チャンスを利用するか否か
- ポイントアップ時に人脈と学内のどちらをアップするか
- 学内昇進するか否か
- 年度末に固定カードを維持するか捨てるか

3.2 プレイヤーエージェントの戦略獲得

本研究では，D-HAL2006 における 7 種のゴールはそれぞれゲーム開始時に決められたものから変わらない，すなわちゴール変更イベントでは常にそれまでと同じゴールを選択するものとして，上記のルールのうち，ゴール以外のルールについて GA による戦略獲得を行った．GA の各個体は以下に示される各部分個体の集合からなる．「順列」「ブール」「リスト」はそれぞれの部分個体の形式であり，D-HAL2006 における行動ルールの仕様に依っている．

- 大学の選好順序（順列）
- 学生の選好順序（順列）
- ポスドクの選好順序（順列）
- ポスドク受入の可否（ブール）
- 年度末学内昇進の可否（ブール）
- 転任チャンスの利用の可否（ブール）
- 初期テーマの選択（リスト）
- 研究テーマの選択（リスト）
- ワイルド時のカードの選択（リスト）
- 任意カードの適用の可否（リスト）
- 教授公募への応募の可否（リスト）
- ポイントアップ時の人脈または学内の選択（リスト）
- 学内昇進カードの適用の可否（リスト）
- 年度末固定カードの取扱方法（リスト）

GA における初期化と突然変異は全ての部分個体に対して行われる．交叉は部分個体の 1 つをランダムに選択し，その部分個体に対して行われる．以下，上記の各形式における初期化，交叉および突然変異の方法を説明する．煩雑さを避けるため，各項においてはここで言う部分個体を単に個体と呼ぶ．

順列型 それぞれの ID を各遺伝子とし，それを並べたものを個体とする．個体中の遺伝子は重複が許されないため，順列として初期化した後，交叉方法として部分写像交叉を用い，突然変異は遺伝子の順序の交換により行う．

ブール型 ルールの仕様に従い，遺伝子として真偽とドントケアをとる．初期化と突然変異はランダムに真偽あるいはドントケアを選択，交叉は交換として行われる．

リスト型 条件が満たされた時に決定を行うサブルール 1 つを 1 個の遺伝子とする．サブルールの個数は可変であるためリスト構造となる．初期化は個体に確率 $p_{\mathrm{next}} = 0.8$ で次のサブルールを連結することで行い，交叉は一点交叉により行う．突然変異はサブルールの追加と削除およびサブルールの位置の交換である．各サブルールの初期化は，200 ほどある条件部それぞれに確率 $p_{\mathrm{val}} = 0.01$ でドントケアでないことを示すランダムな値を設定し，決定部には可能な選択肢からランダムに 1 つ選択することで行う．各サブルールの突然変異は条件部および決定部をランダムに変更する（条件部はドントケアへの変更を含む）ことで行う．

個体間の交叉と突然変異により進化した個体を D-HAL2006 上でシミュレーションにより評価し，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，より早くゴールすることのできる個体（ルールの組合せ）を発見する実験を行った．D-HAL2006 は複数人が同時にプレイするゲームであるが、ここでは対象とする個体それぞれが単独でゲームを行った時にゴールするまでのターン数を求め，個体の評価として利用した．

4. 実験結果

まず，ゲームの性質を調べるために，ランダムに行動選択を行った場合の結果を表 1 に示す．これは，ゴールのみをルールで指定したエージェントによりゲームを 100 試行行った結果である．1 行目はゴールの種類であり，各列はそれぞれをゴールに設定した場合の平均ターン数とその標準偏差，ゴールするまでに要した最小ターン数および，途中で選択可能な研究テーマが無くなった失敗試行数を示している．この結果から，学術社会型と組織研究型はいずれも全試行が失敗しており，ランダムな行動選択ではゴールにたどり着けないことが分かる．一方で，教育者型や悠々自適型はランダムに行動選択する場合でも比較的少ないターン数でゴールし得ることを表しており，比較的簡単であると言える．

次に，3.2 節で定義した個体による単純 GA により，ランダムな初期 100 個体を 100 世代進化させた結果を表 2 に示す．以下で示すのは，進化の過程で出現した最良（すなわち最小ターン数でゴールした）エージェントを，その戦略を固定して 100 試行ゲームを行った結果である．ここでは，選択した 2 個体を交叉する確率を $p_c = 0.9$，着目する遺伝子の突然変異確率を $p_m = 0.01$ とし，前世代の最良個体を次世代に引き継ぐエリート選択を行い，適合度のスケーリングは行わないものとする．最後の行は，用いた最良個体が得られた世代数を示している．ランダム行動では到達不可能だった学術社会型の全試行，および組織研究型の 76 試行ではゴールに到達していることが分かる．また，得られた最小ターン数を比較すると、教育者型や悠々自適型より学内政治型が少なくわずか 28 ターンとなっている．さらに，悠々自適型で 34 ターン，業績卓越型で 35 ターンであり，これらは人間の専門家がゲームを行うときのターン数と比べても遜色が無い．しかし一方で，100 試行の平均ターン数となると，学内政治型で約 70 ターンかかっており，その他でも約 50〜90 ターンと人間に比べて明らかに多くなっている．

次に，突然変異確率の影響を調べるために，突然変異確率を $p_m = 0.1$ に変えた実験を行った．その結果を表 3 に示す．最小ターン数は学術社会型以外で 50 未満と改善されており、平均ターン数でも教育者型，業績量産型，組織研究型が改善されているが，学内政治型のように平均ターン数と失敗数が大きく悪化しているものもあるため、一概に突然変異確率の変更が良いかは決められなかった．

次に，この実験で基準となる表 2 の設定のうち，エリート選択と適合度のスケーリングについて変更した影響を調べた．その実験結果を表 4, 5, 6 に示す．実験結果より，適合度のスケーリングは，この問題において，あまり好ましくない結果になるように思われる．一方で，エリート選択については，有効の場合には平均ターン数が少なくなるが，無効な場合には失敗数がゼロになっているという点が注目される．ここでは進化過程で得られた最良個体の戦略について議論しているため，エリート選択ありの場合には，最良個体がある狭い範囲の条件に過適合していることが考えられる．

表 1: ゲーム中でランダムな行動選択を行った場合のゴールまでの平均ターン数と標準偏差，最小ターン数および失敗数

	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
平均	150.74	130.02	344.83	—	163.02	—	234.89
偏差	33.51	31.69	32.92	—	14.24	—	31.48
最小	87	77	271	—	134	—	167
失敗	46	39	82	100	35	100	43

表 2: 100 個体 100 世代の進化過程で得られた最良個体のルールを固定して 100 試行行った結果．$p_c = 0.9$，$p_m = 0.01$，エリート選択あり，スケーリング無し

	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
平均	90.55	48.96	71.72	93.14	84.68	90.63	60.94
偏差	29.63	8.95	30.44	14.08	11.82	30.05	14.96
最小	44	34	28	54	59	49	35
失敗	0	0	0	0	0	24	0
世代	64	27	23	78	37	35	82

表 3: $p_m = 0.1$ と変更した結果

	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
平均	69.38	52.02	108.71	112.74	69.16	69.59	87.32
偏差	23.27	7.37	121.12	36.79	11.65	15.67	24.67
最小	36	35	25	59	45	41	49
失敗	0	0	27	8	0	0	0
世代	4	67	79	45	53	50	44

さらに，同じ計算量の時の個体数と世代数による違いを調べるために，基準となる表 2 の設定 100 個体 100 世代を，それぞれ 10 個体 1000 世代、20 個体 500 世代，50 個体 200 世代に変更した実験結果を表 7, 8, 9 に示す．その結果，ゴールによって差異が見られるが，表 2 に掲載した，100 世代 100 個体による結果が全体として最良のものとなった．

最後に，人間がプレイした履歴を用いて，行動クローニングで，エージェントを作成した結果 [市瀬 09] と，この実験で得られた結果との比較を行った．行動クローニングによって，作成されたエージェントの平均ゴールターン数は，54.67 から 59.3 であった．一方，表 2 の設定では，ゴールの種類に応じて，平均ゴールターン数は，48.96 から 93.14 であるため，行動クローニングを用いた場合の方が，全般的に良好であると言える．しかし，行動クローニングによるモデルでは，本研究で対象としている「任意カードの適用の可否」のみを取り扱っており，その他の規則に関しては，専門家が予め手動により行動モデルを生成して与えている．行動クローニングにおいては，数少ない事例から，重要なファクターを取り出し，多くの行動に対するモデルを同時に生成することが難しいためである．しかし，本研究で用いたようなシミュレーションを用いると，多くの行動に対するモデルを同時に生成することが可能となるため，これらの研究を統合すると，エージェントの性能を上げることに，大きな役割を果たすことが期待できると考えられる．

5. おわりに

本研究では，人間の行動と同等の行動モデルを学習するために，シミュレーション環境を用いて行動モデルを学習する手法を提案した．その手法は，GA に基づく手法である．HAL2006 というゲーム型学習教材をシミュレーション環境として実験を行った結果，提案手法を用いると，一部の場合において，人間の専門家がゲームを行う場合と，遜色の無いターン数でゴールする行動モデルを生成することができた．

今後は，本研究で取り組んだシミュレーションによる行動モデルの学習と，行動クローニングによる学習の統合化を行い，より適切な行動モデルの学習を行う手法を開発していく予定である．

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)

[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, CD-ROM (2008)

[市瀬 09] 市瀬 龍太郎, 山川 宏：ゲーム型教材における専門家エージェントの考察, 人工知能学会研究会資料, SIG-ALST-A902, pp. 55–60 (2009)

表 4: エリート選択を無効化した結果

	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
平均	99.33	48.93	68.32	115.69	92.07	71.97	117.96
偏差	43.93	8.26	24.21	28.48	13.08	17.37	28.82
最小	34	34	26	52	57	44	59
失敗	0	0	0	0	0	0	0
世代	41	71	24	97	77	82	80

表 5: 適合度のスケーリングを導入した結果

	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
平均	97.57	53.98	101.87	92.56	65.31	79.00	129.37
偏差	55.39	9.10	51.95	22.76	11.15	15.33	33.92
最小	32	34	38	50	43	47	63
失敗	3	0	1	0	0	0	19
世代	47	30	23	69	94	8	41

表 6: エリート選択を無効化し、適合度のスケーリングを導入した結果

	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
平均	112.62	48.58	109.07	93.42	83.51	82.68	111.69
偏差	72.04	7.19	80.94	20.62	14.71	17.92	29.97
最小	47	33	33	61	56	50	56
失敗	21	0	3	0	0	26	1
世代	7	41	62	76	71	45	87

表 7: 10 個体 1000 世代の進化過程で得られた最良個体のルールを固定して 100 試行行った結果．$p_c = 0.9$，$p_m = 0.01$，エリート選択あり，スケーリング無し

	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
平均	107.10	60.53	75.68	94.13	98.87	112.80	70.01
偏差	37.27	11.54	34.63	19.89	15.69	24.48	15.67
最小	48	36	20	46	57	72	44
失敗	3	0	0	0	0	0	0
世代	969	666	971	934	654	740	551

表 8: 20 個体 500 世代の進化過程で得られた最良個体のルールを固定して 100 試行行った結果．$p_c = 0.9$，$p_m = 0.01$，エリート選択あり，スケーリング無し

	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
平均	101.66	53.78	97.82	92.49	80.82	76.46	137.54
偏差	46.68	15.49	62.90	21.76	12.87	24.47	24.77
最小	31	34	38	57	54	45	77
失敗	3	0	2	0	12	1	0
世代	191	426	387	439	188	446	55

表 9: 50 個体 200 世代の進化過程で得られた最良個体のルールを固定して 100 試行行った結果．$p_c = 0.9$，$p_m = 0.01$，エリート選択あり，スケーリング無し

	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
平均	115.21	52.09	112.48	103.17	95.71	76.66	70.59
偏差	46.90	7.42	42.20	23.58	22.98	18.62	20.54
最小	38	37	34	56	64	44	40
失敗	3	0	7	0	0	0	1
世代	139	140	198	75	30	194	185