カードゲームにおける学習エージェントのための状態空間の自動決定

○伊藤崇（青山学院大学），岡松衛，高橋健一（広島市立大学）

Automatic Determination of State Space for Learning Agents in Card Games

* T. Ito (Aoyama Gakuin University), M. Okamatsu and K. Takahashi (Hiroshima City University)

概要ー　ゲームに関する人工知能の研究は歴史が長く，様々なゲームが取り上げられて来た．その中で本研究は，多人数不完全情報ゲームであるトランプゲームのハーツに着目した．不完全情報ゲームとは，相手の手札など直接取得できない情報が存在するゲームのことである．考慮すべき情報が多いため，囲碁などの完全情報ゲームより難しいとされる．これまでに，知覚・行動・報酬を同時にマッピングし学習できる FALCON がハーツに対して有効であることが示されている．しかしながら，先行研究では状態空間の最適化を人手で行っていた．そこで，本研究では FALCON における状態空間の自動決定について検討する．

キーワード: 強化学習，ファジィ ART，FALCON，多人数不完全情報ゲーム，自動決定

1　はじめに

人工知能は広く知られた研究分野であり，Google DeepMind社が開発した囲碁の人工知能「AlphaGo」を皮切りに近年活発に研究が行われている．ゲームに関する人工知能の研究は歴史が長く，手法や計算機の発展によって，チェスからより情報量の多い囲碁へ対象を移してきた歴史がある[1]．現在は囲碁においてもプロ棋士と同等もしくはそれ以上の実力を持った人工知能が開発されており[2]，今後もより複雑なゲームに応用されることが考えられる．これまでに研究されてきたチェスや囲碁は，2人完全情報ゲームと呼ばれ，盤面からゲームに関する全ての情報を所得できるゲームである．実際には制限時間や計算機の問題による制約はあるが，完全情報ゲームでは全ての手を網羅することが可能であり，それによりゲームの展開を完全に予測することができる．人工知能を現実の問題に適用する際に，完全に予測を行うことは不可能であり，予測できない問題も考慮して学習する必要がある．そこで本研究では，より現実問題に近いゲームとして不完全情報ゲームを対象とする．不完全情報ゲームとは，相手の手札などゲームの展開を予測するうえで考慮しなければならない情報が，全て把握できないゲームのことである．不完全情報ゲームは，一般的に完全情報ゲームよりも困難な問題とされている．本研究では，多人数不完全情報ゲームであるトランプゲームのハーツを対象とした．

ハーツを対象とした学習エージェントの研究として，線形関数近似により事前知識を用いない不完全情報ゲームのためのQ学習が提案されている[3]．Q学習は強化学習の中でも一般的な手法の1つであるが，不完全情報ゲームのように考慮すべき情報が多い問題では，状態空間が膨大になってしまう問題がある．また，モンテカルロ学習と多層パーセプトロンを組み合わせた研究[4]も行われている．本研究では，先行研究により多人数不完全情報ゲームに有効性の示されているFALCON（a Fusion Architecture for Learning, COgnition, and Navigation）[5]を用いて，ハーツの学習エージェントを作成する．FALCONは，Ah-wee Tanによって提案されているファジィARTを複数チャネルに拡張した手法であり，知覚・行動・報酬の3つを同時にマッピングし，学習することができる．また，それによりQ学習よりも状態空間を小さくすることが可能となった．さらに先行研究において，不完全な情報の予測に対してSVR（サポートベクター回帰）による予測を組み合わせたFALCONが提案されている[6]．SVRを組み合わせたFALCONでは，行動選択の際に事前に作成されたSVRの判別器により行動の危険度が計算され，ゲームの展開に有利になる行動を選択するように改良されている．

本研究では，FALCONに設定する状態空間の自動決定手法を検討する．先行研究におけるFALCONでは，人手によって最適化された状態空間を用いていた．FALCONを現実問題に応用する場合，状態空間を問題ごとに人手で最適化することは多くの労力と時間を必要とする．そこで，状態空間を自動的に決定する手法の検討を行う．

2　ハーツ

トランプゲームのハーツは，一般的に4人で行われ，ジョーカーを除いた52枚のカードを用いる．ハートやスペードなどの記号をスートと呼び，スートごとにAのカードが最も強くなるように「A, K, Q, … , 4, 3, 2」の順でカードの強さが設定されている．各プレイヤには，ゲームのはじめに13枚のカードがランダムに配布され，自分の番に必ず1枚のカードをプレイする．全員が1枚ずつカードをプレイし，場のカードで勝敗が決まるまでをトリックと呼び，13トリック連続して行うことで1ゲームが終了する．また，各トリックで最初にプレイされるカードをリーディングカードと呼び，それをプレイしたプレイヤを親と呼ぶ．また，それ以外のプレイヤを子と呼ぶ．子は，リーディングカードと同じスート（リーディングスート）のカードをプレイしなければならない．ただし，同じスートのカードがない場合に限り，任意のカードを出すことができる．全員がカードをプレイした後，リーディングスートで最も強いカードをプレイしたプレイヤが場の全てのカードを引き取り，次のトリックの親となる．そして，13トリック終了後，各プレイヤの引き取ったカードから罰点の集計が行われ，ハートは1枚1点，スペードのQは1枚13点として計算される．ハーツにおける各プレイヤの目的は，この罰点を少なくすることである．

基本的なルールとして,1トリック目の親はクラブの2を持っているプレイヤとし,そのプレイヤはクラブの2をプレイしなければならない．また，1トリック目には罰点のカードは優先してプレイできない，ハートブレイク状態になるまで罰点のカードはリーディングカードにできないといったルールがある．ハートブレイクは戦略的に重要な要素であり，いずれかのプレイヤによりハートのカードがプレイされた，もしくはスペードのQがプレイされた場合，ハートブレイク状態になる．

その他のルールとして，手札交換やシュートザムーンなどがあるが，これらのルールは採用しない．

3 FALCON

3.1 アルゴリズム

FALCON[5]はファジィ ART を複数チャネルに拡張した手法であり，知覚・行動・報酬の 3 つのチャネルで構成されている．状態空間（SF），行動空間（MF），報酬空間（FF）はそれぞれがカテゴリ（CF）に関連付けられている．エージェントが M 個の知覚センサを有するとき，SF には知覚ベクトル $S=(s_1,\cdots,s_M)$ が与えられる．ここで，s_i は各知覚センサの要素を表すニューロンであり，$[0,1]$ の範囲の値をとる．また，エージェントにより実行可能な行動が K 個ある時，MF には行動ベクトル $A=(a_1,\cdots,a_K)$ が与えられる．ここで，a_i は各行動の要素を表すニューロンであり，$[0,1]$ の範囲の値をとる．FF には，報酬ベクトル $R=(r,1-r)$ が与えられる．r はエージェントが環境から受け取る報酬を表すニューロンであり，$r\in[0,1]$ である．カテゴリは L 個のニューロン $n_j(j=1,\cdots,L)$ を有し，SF, MF および FF のニューロンとそれぞれ重みベクトル $W_j^S=(w_j^{s1},\cdots,w_j^{sM})$，$W_j^A=(w_j^{a1},\cdots,w_j^{aK})$ および $W_j^R=(w_j^{r1},w_j^{r2})$ によって関連付けられている．重みベクトルの要素 $w_j^{yx}\in[0,1]$ は，ニューロン n_j と SF, MF および FF のニューロンとの関連の強さを表す．FALCON は行動選択フェーズと学習フェーズを繰り返し，重みベクトルを更新することによって知覚，行動，報酬の関連を学習して行く．

はじめに，SF に関連付けられている W_j^S と AF に関連付けられている W_j^A を 1，FF と関連付けられている W_j^R を $(1,0)$ に初期化する．

次に行動選択フェーズを行う．エージェントの入力により知覚ベクトル S を取得し，行動ベクトル A の K 個の各ニューロンに 1 を，報酬ベクトルに $(1,0)$ を設定する．そしてそれらを用いて式(1)によりカテゴリごとに選択強度 T を計算する．

$$T_j=\gamma_1\frac{\|S\wedge W_j^S\|}{\alpha+\|W_j^S\|}+\gamma_2\frac{\|A\wedge W_j^A\|}{\alpha+\|W_j^A\|}+\gamma_3\frac{\|R\wedge W_j^R\|}{\alpha+\|W_j^R\|} \quad (1)$$

ここで，二項演算子 $\wedge$ はファジイ AND オペレータであり，$\|\bullet\|$ は各要素の絶対値和である．また，パラメータ γ は非負の実数値であり，本研究では $\gamma=(1.0,0.0,0.01)$ としている．α は分母が 0 になることを防ぐために設定されたパラメータであり,0.001 とする．選択強度が最も高くなったカテゴリを選択し，そのカテゴリに関連付けられている行動ベクトルの中で最もニューロンが大きな行動を実行する．ハーツのための FALCON では 1 ゲームの結果を踏まえて学習を行う．

13 トリック終了後，学習フェーズに移る．学習フェーズでは，ゲームの結果からトリックごとに重みベクトルの更新もしくは新たなカテゴリの追加を行う．まず，罰点を確認し，6 点を超えていれば負のフィードバック，それ以下なら正のフィードバックとする．次に，1 トリック目の知覚ベクトル S，行動ベクトル A を取得する．ここで，負のフィードバックの場合は，そのトリックで行動したニューロン a を 0，それ以外を 1 とし，報酬ベクトルは $(0,1)$ とする．正のフィードバックの場合は，そのトリックで行動したニューロン a を 1，それ以外を 0 とし，報酬ベクトルは $(1,0)$ とする．そして，そのトリックで最も高い選択強度となったカテゴリの重みベクトル W と知覚・行動・報酬ベクトルを,それぞれ式(2), 式(3), 式(4)によって比較する．

$$\frac{\|S\wedge W_j^S\|}{\|S\|}\geq\rho_1 \quad (2)$$

$$\frac{\|A\wedge W_j^A\|}{\|A\|}\geq\rho_2 \quad (3)$$

$$\frac{\|R\wedge W_j^R\|}{\|R\|}\geq\rho_3 \quad (4)$$

ここで ρ は警戒パラメータであり，$\rho=(1.0,0.0,0.5)$ とした．式(2), 式(3), 式(4)をすべて満たすカテゴリであれば, そのカテゴリに関連する重みベクトル W を更新し，いずれかが満たさなければ次に選択強度の高いカテゴリで比較を行う．知覚に関する重みベクトル W^S は式(5), 報酬に関する重みベクトル W^R は式(6)により更新される．式(6)の β は 0.001 としている．

$$W_j^S=(S\wedge W_j^S) \quad (5)$$

$$W_j^R=(1-\beta)W_j^R+\beta(R\wedge W_j^R) \quad (6)$$

行動に関する重みベクトル W^A は 3 つの式により更新される．まず式(7)により学習スピード ε が計算される．*Total_Gnum* は最大ゲーム数，*Gnum* は現在のゲーム数である．次に，そのトリックで罰点があった場合は式(8)により W^A を更新する．最後に，そのゲームのフィードバックが正であれば式(9)により W^A を更新する．

$$\varepsilon=0.00001\times\frac{Total\_Gnum-Gnum}{Total\_Gnum} \quad (7)$$

$$W_j^A=W_j^A-BAD\times\varepsilon \quad (8)$$

$$W_j^A=W_j^A-7.0\times\varepsilon \quad (9)$$

ここで，*BAD* はそのトリックでとった罰点である．報酬に関する重みベクトル W^R ついては更新されない．もし，全てのカテゴリにおいて式(2), 式(3), 式(4)のいずれかが満たされない場合，新しいカテゴリとして追加する．ここで，知覚に関する重みベクトル W^S と報酬に関する重みベクトル W^R は，式(5)と式(6)により設

定される．行動に関する重みベクトルW^Aは全て 1 とする．これらを 13 トリックまで行い，学習フェーズを終了する．

3.2 SVR による危険度予測を取り入れた FALCON

SVR による危険予測を取り入れた FALCON[6]では，行動選択フェーズの中で，選択された行動の危険度が計算される．

まず，通常の FALCON と同様に式(1)を用いて選択強度が最大となるカテゴリを選択する．次に，そのカテゴリに関連付けられている行動ベクトルの中で最もニューロンが大きな行動を選択する．ここで，その行動を予め作成した SVR の判別器に入力し，危険度を計算する．危険度がしきい値より低ければその行動を行い，そうでなければ次にニューロンの大きな行動を選択し，危険度の計算を行う．全ての行動の危険度が条件を満たさなければ，最もニューロンの大きな行動を行う．

4 状態空間の自動決定手法

4.1 状態空間

本研究で探索した状態空間を Table 1 と Table 2 に示す．Table 内の S はスペード，H はハート，D はダイヤ，C はクラブ，L はリードスートのことである．また，SJ，SQ，SK，SA はそれぞれスペードの J，Q，K，A を示す．知覚ベクトルは 0 もしくは 1 で表現され，真であれば 1 をとる．Table 1 の No.7 から No.10 と Table 2 の No.4 は，4 つの取りうる状態を 2 ビットで表現している．

先行研究では，この中から人手で最適な知覚情報の組み合わせを見つけ出していた．先行研究で使用されていた知覚情報は，親番では No.6,No.14,No.15，子番では No.0,No.3,No.9,No.10,No.11 である．

Table 1　親番における知覚情報

No.	知覚情報	ビット数
0	ハートブレイク状態	1
1	・手札の中で最も枚数の多いスートが D ・手札の中で最も枚数の多いスートが C	2
2	H に関する状態： ・手札に H が 1 枚 ・手札の最強の H より強い H を敵が所持 ・手札の最強の H より弱い H を敵が所持 ・手札の最弱の H より強い H を敵が所持 ・手札の最弱の H より弱い H を敵が所持	5
3	S に関する状態： No.2 と同様	5
4	D に関する状態： No.2 と同様	5
5	C に関する状態： No.2 と同様	5
6	SQ/SK/SA に関する状態 ・SQ を自分が所持 ・SQ を敵が所持 ・SK を自分が所持 ・SK を敵が所持 ・SA を自分が所持 ・SA を敵が所持	6
7	C に関する枚数： ・自分の手札の C の枚数（0, 1~2, 3~5, 6 枚以上） ・相手の手札の C の枚数（0, 1, 2, 3 枚以上）	4
8	D に関する枚数： No.7 と同様	4
9	S に関する枚数： No.7 と同様	4
10	H に関する枚数： No.7 と同様	4
11	各スートについて未所持の敵がいる	4
12	各スートについて自分が所持	4
13	各スートについて自分が平均以上のカードを所持	4
14	自分が SJ 以下を所持	1
15	相手が SJ 以下を未所持である	1

4.2 状態空間の自動決定

本研究で提案する状態空間の自動決定では，親番と子番それぞれに対して適した知覚情報の組み合わせを探索する．以下に状態空間の自動決定手法のアルゴリズムを示す．

自動決定のアルゴリズム

① 親番であれば Table1，子番であれば Table2 に示す知覚情報を知覚情報の候補群として設定し，最良評価値を 1.0 とする．

② 候補群の中から知覚情報を 1 つ選択する．

③ 登録されている知覚情報と②で選択された知覚情報を与えた FALCON と対戦エージェント 3 体を 75,000 ゲーム対戦させ，学習を行う．

④ 学習した FALCON を用いて対戦用エージェン

Table 2　子番における知覚情報

No.	知覚情報	ビット数
0	ハートブレイク状態	1
1	・手札の中で最も枚数の多いスートが D ・手札の中で最も枚数の多いスートが C	2
2	L に関する状態： ・手札に L が 1 枚しかない ・手札の最強の L より強い L を敵が所持 ・手札の最強の L より弱い L を敵が所持 ・手札の最弱の L より強い L を敵が所持 ・手札の最弱の L より弱い L を敵が所持	5
3	SQ/SK/SA に関する状態 ・SQ を自分が所持 ・SQ を敵が所持 ・SK を自分が所持 ・SK を敵が所持 ・SA を自分が所持 ・SA を敵が所持	6
4	L に関する枚数： ・自分の手札の L の枚数（0, 1~2, 3~5, 6 枚以上） ・相手の手札の L の枚数（0, 1, 2, 3 枚以上）	4
5	自分より後に L を未所持プレイヤがいる	1
6	各スートについて自分が平均以上のカードを所持している	4
7	自分が SJ 以下を所持している	1
8	相手が SJ 以下を未所持	1
9	L のスート	2
10	自分の所持している L の枚数　(0, 1, 2 枚以上)	3
11	手札の L の最強カードが残っているもので最も強い	1
12	場にペナルティーカードがある	1

ト 3 体と 5,000 ゲーム対戦させ，平均獲得罰点比率を算出する．

⑤ ③~④を 10 試行行い，平均獲得罰点比率の平均を評価値として算出する．

⑥ ② ~⑤までを全ての知覚情報の候補に対して実行する．

⑦ 最良評価値と⑥の中で最も低い評価値を比較し，最良評価値を下回らなければここで探索を終了する．

⑧ 最も低い評価値を示した知覚情報の候補を問題に適した知覚情報として登録し，候補群から削除する．また，その評価値を最良評価値として保存する．

⑨ 知覚情報の候補が残っていれば，②から探索を再開する．

ここで評価値として採用している平均獲得罰点比率は，獲得した罰点の合計を対戦ゲーム数×1 ゲームの最大罰点 26 で割った値である．したがって，0.25 を下回るとそのプレイヤは他のエージェントよりも強いといえる．

自動決定手法では，親番に対する状態空間の探索と子番に対する状態空間の探索をそれぞれ独立して行う．これは，親番と子番での別々の FALCON を用いているためである．そこで，親番の状態空間を探索する際には，子番の行動選択はモンテカルロエージェントに行わせ，子番の状態空間を探索する際には，親番の行動選択はモンテカルロエージェントに行わせる．また，自動決定のアルゴリズム ②の対戦エージェントは，前半の 37,500 ゲームではルールベースエージェント，後半の 37,500 ゲームではモンテカルロエージェントとした．ルールベースエージェントとモンテカルロエージェントについては，5.1 節で後述する．

5 実験

5.1 対戦エージェント

本研究では，対戦エージェントとしてルールベースエージェント，モンテカルロエージェントの 2 種類のエージェントを用いた．

ルールベースエージェントは，gnome-hearts を元にしたルールを搭載しているエージェントである．自分の所持しているカードの種類や場の状態，未提出のカードの情報などを考慮し，設定された if - then ルールに基づきプレイするカードを決定する．

モンテカルロエージェントは，UCT モンテカルロ法を用いてゲームの状態をシミュレートし行動を選択するエージェントである．現在のトリックをシミュレート対象とし，残っているカードをランダムに相手プレイヤに振り分け，ゲーム終了までのシミュレーションを 200 回行う．その中で，最も獲得罰点が少なくなった行動を選択する．

事前に行った実験により，これら 2 種類のエージェントの中ではモンテカルロエージェントが最も強く，次にルールベースエージェントが強いということが分かっている．

5.2 検証実験

検証実験により選択された知覚情報と，対戦エージェントとの対戦結果を Table 3，Table 4 に示す．Table 3 は親番に対して自動決定を行った結果，Table 4 は子番に対して自動決定を行った結果を示し，対戦は平均獲得罰点比率としている．Table 3 から，親番に対しては先行研究で最適とされていた(6),(14),(15)を含む 5 つの知覚情報が選択されている．Table 1 の各知覚情報の内容を見ると，(6),(14),(15)はどれもスペードに関する知覚情報であり，追加で選択された(3),(9)についてもスペードに関する知覚情報であった．したがって，人手で選択したものと同等の状態空間が設定できたと考えられる．また，対戦結果から先行研究とほぼ同等の成績が得られているため，性能で見ても確かなものが選択できていることが確認された．Table 4 から，子番に対しては異なる親番と異なり，人手で選ばれた結果と同じ知覚情報は 2 つだけとなっている．Table 2 より，共通している(0)と(10)は，ハートブレイクの状態とリードカードの所持枚数に関する知覚情報であることがわかる．自動決定で新たに選択された(2)，(5)はリードカードの相手の所持情報を含むものとなっている．その他先行研究で選ばれている(3)，(9)，(11)は，(3)がスペードの Q 以上のカードに関する知覚情報，(9)と(11)がリーディングスートとリーディングカードに関する知覚情報となっている．対戦結果を見ると平均獲得罰点比率に大きな差は見られないため，子番においてはスペードのカードに関する知覚情報は重要度が低いと考えられる．

6 おわりに

本研究では，多人数不完全情報ゲームであるトランプゲームのハーツを対象とした FALCON の状態空間の自動決定について検討を行った．自動決定の検証実験では，先行研究において人手で最適化された状態空間と同様の傾向を持った状態空間が設定され，対戦成績で見ても同等の強さを示している．しかしながら，本研究の方法では多く時間を必要とする点や，知覚情報の順番に選択するため考慮できていない組み合わせが存在するなどの問題がある．そこで，効率的な自動

Table 3 親番における自動決定の結果

	自動決定手法	先行研究
知覚情報	(3), (6),(9),(14), (15)	(6),(14),(15)
モンテカルロ	0. 2547	0.2500
ルールベース	0. 2174	0.2173

Table 4 子番における自動決定の結果

	自動決定手法	先行研究
知覚情報	(0), (2), (5),(10)	(0),(3),(9),(10),(11)
モンテカルロ	0.2547	0.2500
ルールベース	0.2191	0.2173

決定の方法を検討する必要がある.

謝辞

本研究は AOYAMA VISION「AI 研究拠点形成プロジェクト」による支援を受けた.

参考文献

1) 斉藤 康己：コンピュータ囲碁研究（<小特集>「ゲームプログラミング」）, 人工知能学会誌, **10**-6, 860/870（1995）
2) 加藤 英樹 : Zen のアーキテクチャ（<特集>コンピュータ囲碁）, 人工知能学会誌, **27**-5, 501/504（2012）
3) 齋藤 雄太, 鶴岡 慶雅：線形関数近似によるトリックテイキングゲームの Q 学習, ゲームプログラミングワークショップ 2016 論文集, **2016**, 196/200（2016）
4) M. Wagenaar：Learning to play the Game of Hearts using Reinforcement Learning and a Multi-Layer Perceptron, Diss. Faculty of Science and Engineering（2017）
5) A. H. Tan：FALCON: A Fusion Architecture for Learning, COgnition, and Navigation, International Joint Conference on Neural Networks, **4**, 3297/3302（2004）
6) 笠原 和真, 二本 健太, 伊藤 崇, 高橋 健一, 稲葉 通将 : SVR を適用した FALCON によるトランプゲームに対する学習実験, 日本知能情報ファジィ学会誌 : 知能と情報, **30**-4, 643/651（2018）