マルチエージェントを用いたチームAIと個別AIについて

日大生産工(院)　○藤田　真広
日大生産工　　齋藤　敏雄

1. まえがき

近年の人工知能技術の発展はめまぐるしいものがあり，様々な分野に応用されてきた．

ゲーム業界においては，特にグラフィックに偏り発展してきた歴史と，メモリ容量問題や人工知能技術に計算時間が必要なものが多いなどの理由により，広く利用されているとはいいがたく，ごく一部の分野において利用されていたに過ぎない．しかし，ハードウェアの進化により，シミュレーション能力が向上し，ゲーム基本システムに対する余剰リソースが発生した．これにより，グラフィックのみならず，人工知能技術をゲームに組み込むことが可能となった．

本研究では，“スコットランドヤード”というボードゲームを題材として，AIを備えたエージェントをプレイヤーとして定式化し，チームとしての振る舞いと個人としての振る舞いを比較検討する．結果として，ゲームに人工知能技術を応用することで，ゲーム製作に対し新しい幅を提示する事を目的とする．

2. スコットランドヤード

スコットランドヤードは，ドイツのランベスバーガー社より発売された代表的なボードゲームである．ロンドン市内をモチーフにしており，4～5人（本研究では5人を想定）の刑事と1人の怪盗にわかれ，盤上の1～199の地点をそれぞれ，タクシー・バス・地下鉄・船（怪盗のみ）を使い移動する．また，移動手段はそれぞれ使用回数を制限されている．刑事側は怪盗を捕まえるか，移動できないように追い詰めることが勝利条件で，怪盗側は24時間逃げ切れば勝ちとなる．

3. マルチエージェントシステム

3.1 マルチエージェントシステム[1)2)]

マルチエージェントシステムは，自立的に行動する多数のエージェントから構成される．それぞれのエージェントは自分が置かれている環境を知覚し，自分の目標を達成するように行動を選択する．エージェントは互いに影響を及ぼしあい，それが，各エージェントの行動選択基準を変化させるキッカケにもなる．マルチエージェントシステムには，協力型マルチエージェントシステムと競争型マルチエージェントシステムがあるが，本研究ではその両方を用いるものとする．

3.2 強化学習[3)4)]

強化学習は教師なし学習のひとつであり，環境の状態sに対して行動aをとったときに環境から得られる報酬rをもとに，初期状態からゴール状態に渡って受け取る報酬が最大になるような行動戦略を学習する．(Fig.1)
環境に関する正しい知識をあらかじめ準備する必要がなく，行動とその行動の評価を繰り返しながら学習していくため，環境が変化する動的な環境にも対応することができる．

Behavior of Multiagent System with Team AI

Masahiro FUJITA and Toshio SAITO

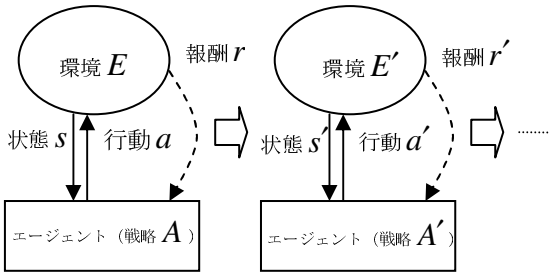

Fig.1 強化学習における相互作用

3.3 マルコフ性

マルコフ性とは，確率論における確率過程の持つ特性の一種で，その過程の将来状態の条件付き確率分布が，現在状態のみに依存し，過去のいかなる状態にも依存しない特性を持つことをいう．本研究では，エージェントの現在位置と残りの移動手段をマルコフ状態として扱う．よって，$t+1$における環境の応答はtにおける状態と行動表現のみに依存することになり，このときには全てのs'，r，s_tとa_tに対して

$$\Pr\{s_{t+1}=s', r_{t+1}=r \mid s_t, a_t\} \qquad (1)$$

のみを指定することで環境のダイナミクスを定義することができる．

3.4 有限マルコフ決定過程（有限 MDP）

マルコフ性を満たす強化学習タスクはマルコフ決定過程（MDP）と呼ばれる．また，本研究では状態と行動の空間が有限であるので，有限 MDP であるといえる．有限 MDP は状態と行動の集合と，環境の 1 ステップダイナミクスから定義される．次に可能な各状態s'の確率は，

$$P_{ss'}^{a}=\Pr\{s_{t+1}=s' \mid s_t=s, a_t=a\} \qquad (2)$$

これらの量は遷移確率と呼ばれている．同様にして，次の報酬の期待値は，

$$R_{ss'}^{a}=E\{r_{t+1} \mid s_t=s, a_t=a, s_{t+1}=s'\} \qquad (3)$$

3.5 価値関数

価値関数は状態の関数で，エージェントがある状態にいることがどれだけ良いのかを評価する．ここでは，どれだけ良いのかという概念を将来において期待される報酬に関して定義する．エージェントが将来受け取ることを期待できる報酬は，エージェントがどのような行動を取るかに依存する．したがって，価値関数は特定の方策に関して定義される．

方策πが各状態$s\in S$と行動$a\in A(s)$から，状態sで行動aを取る確立$\pi(s,a)$への写像であるといえる．その時，MDP に対する

$$V^{\pi}(s)=E_{\pi}\{R_t \mid s_t=s\}\}$$

$$=E_{\pi}\{\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1} \mid s_t=s \qquad (4)$$

$E_{\pi}\{\}$は，エージェントがπに従うとしたときの期待値を表す．関数V^{π}を方策πに対する状態価値関数と呼ぶ．

同様に，方策πのもとで，状態sにおいて行動aを取る事の価値を$Q^{\pi}(s,a)$で表し，状態sで行動aを取り，その後に方策πに従った期待報酬として定義する．

$$Q^{\pi}(s,a)=E_{\pi}\{R_t \mid s_t=s, a_t=a\}$$

$$=E_{\pi}\{\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1} \mid s_t=s, a_t=a\} \qquad (5)$$

Q^{π}を方策πに対する行動価値関数と呼ぶ．

3.6 最適価値関数

有限 MDP に対しては，以下のようにして最適方策を定義することができる．価値関数は方策に関して反順序を定義する．すべての状態に対して，方策πの期待収益がπ'よりも良いか同じであるなら，πはπ'よりも良いか、

同じであると定義される．言い換えるなら、すべての$s \in S$に対して，$V^{\pi}(s) \geq V^{\pi'}(s)$であるなら，そのときに限り$\pi \geq \pi'$である．他の方策よりも良いか，それに等しい方策が常に少なくとも 1 つ以上存在し，これが 1 つの最適方策である．最適方策は 1 つ以上存在するかもしれないが，全ての方策をπ^*と記す．最適方策群は，最適状態価値関数と呼ばれる，同じ状態価値関数を共有する．最適状態関数は，すべての$s \in S$に対して

$$V^*(s) = \max_{\pi} V^{\pi}(s) \qquad (6)$$

と定義される．

4. スコットランドヤードにおけるマルチエージェントの実装

4.1 移動方法と移動可能範囲

怪盗は 3・8・13・18・24 時間目以外は姿が見えないが，毎時間どのように移動したかは刑事にもわかる．移動にはチケットを利用し，最初に刑事には TAXI チケット 10 枚，バスチケット 8 枚，地下鉄チケット 4 枚を渡され，これを用いて移動する．怪盗は TAXI・バス・地下鉄は無制限に移動でき，それとは別に，ブラックチケット 5 枚とダブルムーブ 2 枚が渡される．ブラックチケットは TAXI・バス・地下鉄の他に，船を使って移動する事ができる．また，ブラックチケットを使用したときは，刑事には移動手段がわからない．ダブルムーブはその名の通り，2 時間連続で移動することが可能になる．

199 の移動地点（Fig.2）は，すべてある別の移動地点より TAXI で移動可能であり，バスを利用可能な地点は 59 地点，地下鉄が利用可能な地点は 13 地点である．また，船で移動できる地点は 4 地点である．

Fig.2 スコットランドヤードの地図とチケット 5)

4.2 刑事の行動

刑事の移動は 5 人のエージェントによる協力型マルチエージェントシステムであるとして、怪盗を逮捕するように動く．

4.2.1 協力型マルチエージェントシステムの利用

刑事は，怪盗を逮捕するという問題を，直接逮捕に向かう刑事，地下鉄や船を怪盗に利用させないために駅などを押さえる刑事，全体的なゾーンで移動範囲を狭める刑事という副問題へと割り当て，分割する．各エージェントはマルコフ性を持ち合わせているために，この副問題の割り当ては永続的なものではなく，毎時間変化するものとする．

4.2.2 移動手段の選択

3 時間目までは，怪盗の位置がわからないために地下鉄や船を利用させないように動く．3 時間目以降は，それぞれに副問題を割り当て移動させる．怪盗までの距離・怪盗の移動手段による存在しない地域への可能性の除去・残っている移動手段・現在地などを評価することで副問題をどのエージェントが担当するかを割り当てる．その割り当てと残りの移動手段をもとに，移動手段を選択する．また，思考の範囲は怪盗が姿を現した時間から，次に姿を現す時間までの範囲で思考する．

4.3　怪盗の行動

全体のシステムとして，怪盗と刑事は競争型マルチエージェントであるとして，刑事から逃げる．

現在から 5 時間後までの刑事の移動可能範囲を考える．1 時間で移動可能な場所を報酬 $r=-1$ として，そこから，1 時間ごとに r を割引きする．また，複数の刑事が到達可能な場所は，r 同士を加算する．また，移動可能な手段の種類と量により一定の r_p を加算する．そうして，得られた報酬のうち高いものを選択し，移動する．得られる最大の報酬がある一定以下になった場合には，ブラックチケットやダブルムーブを使用する．また，姿を現した直後からブラックチケットを使うまでは一定の報酬を減算する．これは，刑事側に現在位置を推測されにくくするためである．

5. 結果と考察

本研究では，100 回の試行を 1 セットとし，100 セット行う．1 セット毎に勝敗により報酬の値を変化させ，刑事・怪盗ともに学習させていく．1 セット毎の勝率の推移と，刑事が逮捕した場合にかかった時間の推移を比較し，正しく学習できたか確認する．正しく学習できている場合は，刑事・怪盗ともに最適方策が収束していくために，勝率は一定の確率に収束していき，逮捕した場合もかかる時間は増加していくはずである．また，怪盗よりも刑事の方が報酬が曖昧であるために，収束に時間が掛かることが予想される．

また，刑事を協力させずに，個別の行動に対して，報酬を与え学習させる．この場合でも，100 回の試行を 100 セット行い，1 セット毎の勝敗により報酬の値を変化させる．協力した場合と，しなかった場合の勝率の推移と，逮捕するまでにかかった時間を比較する．

6. まとめ

本研究では，ボードゲームという，エージェントが非常に限られた行動群の中からしか行動を選択しなかったが，行動の幅の広いアクションやシューティングにおいても行動群を一般化することで，マルチエージェントシステムを導入する事が可能であると考える．

また，プランナーやデザイナーの経験則によって学習させるのではなく，繰り返し試行する事によって最適方策を模索するものなので，製作の負担を軽減することが可能であると考える．しかし，経験則による学習ではないために，AI が思考のループに陥ったり，デバッグ時にデバッグ項目を挙げにくいなど，いくつかの問題点を内包している．

今後の展望としては，まだ未完成であるプログラムを完成させ，これらの理論を実証し検証すべきである．また，刑事が，自分たちのいる場所から遠ざかるであろうという推測を報酬に追加した場合と，怪盗が，刑事のいる場所から遠ざかるであろうという推測を一定確率で裏切るという可能性を追加した場合についても比較・検証する必要がある．

参考文献

1) 大内東，山本雅人，川村秀憲，マルチエージェントシステムの基礎と応用，コロナ社，(2002)，pp. 1-90.

2) 高玉圭樹，マルチエージェント学習，コロナ社，(2003)，pp. 21-64.

3) 三上貞芳，皆川雅章，強化学習，森北出版，(2000)，pp. 2-170.

4) 電気学会，学習とそのアルゴリズム，森北出版，(2002)，pp155-179.

5) Rabensburger, http://www.ravensburger.de/web/

6) 三宅陽一郎，人工知能が拓くオンラインゲームの可能性，AOGC2007，(2007)．