情報処理学会論文誌 Vol. 50 No. 12 2796–2806 (Dec. 2009)

戦略型トレーディングカードゲームのための戦略獲得手法

藤井叙人[†1] 片寄晴弘[†1,†2]

市販テレビゲームにおいて，ゲーム内のコンピュータ（COM）の戦略に対してプレイヤの意識が高まりつつある．特に，世界的に人気のある遊☆戯☆王やポケットモンスターに代表されるビデオトレーディングカードゲーム（ビデオ TCG）においては，プレイヤの要求に合わせた COM の強さの設定が必要不可欠である．現在ではゲームプログラマによる戦略の作り込みによって実現されているが，これは非常に煩雑で時間がかかる．本研究では，強化学習法を用いて，戦略型ビデオ TCG の戦略を自動学習する戦略学習機構について検討する．COM の最適行動学習だけでなく，TCG 特有の要素である “最適なカード組合せ” や “魔法や罠などの特殊効果” に関しても学習機構を実装する．戦略学習機構の評価として，ルールベース戦略を相手とした計算機実験を実施する．最後に，戦略学習機構の汎用性と，残された課題，その解決策について検討する．

Strategy-acquisition System for Video Trading Card Game

Nobuto Fujii[†1] and Haruhiro Katayose[†1,†2]

Behavior and strategy of computers (COM) have recently attracted considerable attention with regards to video games, with the development of hardware and the spread of entertainment on the Internet. Previous studies have reported strategy-acquisition schemes for board games and fighting games. However, there have been few studies dealing with the scheme applicable for video Trading Card Games (video TCG). We present an automatic strategy-acquisition system for video TCGs. The proposed strategy-acquisition system uses a sampling technique, Action predictor, and State value function for obtaining rational strategy from many unobservable variables in a large state space. Computer simulations, where our agent played against a Rule-based agent, showed that a COM with the proposed strategy-acquisition system becomes stronger and more adaptable against an opponent’s strategy.

1. はじめに

トレーディングカードゲーム（TCG）は，現在のアミューズメント系ゲームの中でもきわめて人気が高い．世界一売れている TCG の遊☆戯☆王[1] では，2006 年 11 月までの集計売上が 158 億万枚に達しているという報告がある[2]．現在では，遊☆戯☆王は，TCG だけでなく市販テレビゲーム（ビデオ TCG）としても発売されている．また，世界一売れたロールプレイングゲームであるポケットモンスター[3]（ポケモン，2008 年 8 月時点で 1 億 8 千万本売上げ）も，TCG に移植されており，非常に人気のあるタイトルになっている．

ビデオ TCG の大きな楽しみの 1 つとして，コンピュータ（COM）と対戦し，COM に勝つための戦略を組み立てるということがある．プレイヤの手応え感を確保するためには，プレイヤのレベルに応じた COM の戦略を複数用意しておき，プレイヤの上達に応じてその戦略レベルを切り替えるというような機構が必要になる．これまでの市販テレビゲームにおいては，ゲームプログラマによる綿密な作り込みと数多くのデバッグプレイにより，その品質が確保されてきた．しかし，プレイヤのプレイスタイルに COM の戦略を適応させることが可能となれば，制作者にとっては開発効率の向上，プレイヤにとっては楽しみの持続というメリットが生じる．COM の振舞い[4),5)] や，ゲーム戦略の自動学習[6)–10)] に関する研究がさかんに行われるようになってきているが，ビデオ TCG において，プレイヤのプレイスタイルに COM の戦略を適応させる研究はほとんど報告がなされていない．

本研究では，藤田らが提案したサンプリング手法や戦略学習手法[7)–10)] に基づき，戦略型ビデオ TCG における戦略を自動学習する戦略学習機構について検討する．藤田らは，トランプゲームの “Hearts” を対象とし，部分観測状況に起因する巨大な状態空間の問題を解決したうえで，人間の熟達者よりも優れた戦略を得ることに成功している．しかし，戦略型ビデオ TCG を扱う場合には，部分観測状況に起因する巨大な状態空間の問題に加えて，カードを準備する段階でのカードの組合せ,「魔法」などの特殊効果,「罠」のようなある特定の条件を満たすことで発動する効果など，TCG に欠かせないゲームの要素も考慮する必要がある．本研究の戦略学習機構は，あらゆる TCG に適応できる戦略獲得機構の実現を目指す．

以下，2 章で関連研究を紹介し，3 章で，本研究で扱う戦略型ビデオ TCG のルールを設

†1 関西学院大学大学院理工学研究科
Graduate School of Science and Technology, Kwansei Gakuin University
†2 科学技術振興機構戦略的創造研究推進事業 CrestMuse プロジェクト
CrestMuse Project, CREST, JST

定し，戦略学習機構を実装するうえでの困難性を整理する．4 章で，戦略学習機構の実装方法について述べ，5 章で，計算機実験により戦略学習機構の戦略を評価する．6 章で，戦略学習機構の汎用性と，残された課題，その解決策について検討し，7 章で，戦略学習機構の可能性と今後の課題について述べる．

2. 関 連 研 究

2.1 COM の振舞い

ゲームの面白さ向上に，COM の振舞いからアプローチした研究事例として，対戦型ゲームをプレイしているときの脳活動の計測[4]，対戦型ゲームにおいて戦況に応じた感情的な発話をする仮想プレイヤ[5] などがあげられる．

玉越らは，対戦型ゲームをプレイしているときの脳活動を，対人間（対戦相手が互いに観察可能），対人間（対戦相手が人間か COM か観察できない），対 COM の 3 つの条件で計測し，質問紙による主観的評価を実施している[4]．実験の結果から，「操作のしやすさ」と「相手のレベルとの拮抗」がゲームプレイ時の没入感に大きく影響すること，対人間のときに最も没入感を感じやすいことが示されている．

塩入らは，対戦型ゲームにおいて戦況に応じた感情的な発話をする仮想プレイヤを提案している[5]．ゲーム映像からゲーム状況を読み取り，プレイヤからの干渉や環境変化に対して仮想プレイヤが発話する．仮想プレイヤの心理状態を 2 つのパラメータによって制御し，感情的な発話をさせることで，プレイヤはゲームをより面白く感じることができると考察している．

2.2 COM の戦略

ゲームの面白さ向上に，COM の戦略の面からアプローチした研究事例として，将棋[6] などのボードゲームや，トランプゲーム[7)–10)] への戦略学習機構の実装があげられる．これらの研究では，実際の対戦データから COM の戦略を学習させ，相手の戦略に対して臨機応変に行動を出力するような戦略学習機構の実現が目的とされている．

保木は，将棋を対象とした戦略学習機構（コンピュータ将棋プログラム）として Bonanza を提案している[6]．従来のコンピュータ将棋プログラムは，評価関数や選択的探索を人手で経験的に決定していた．しかし，Bonanza は，プロ棋士の棋譜 6 万局から評価関数を自動学習する手法と，ある局面においてすべての可能性を考慮する全幅探索手法を採用することで，従来手法よりも良い戦略を得ることに成功し，コンピュータ将棋界に大きな衝撃を与えている．保木は，自動学習による将棋の戦略学習機構は，10〜20 年後にはプロ棋士にも勝る戦略を得ることができるだろうと考察している．

藤田らは，トランプゲームの “Hearts” を対象とした COM の戦略学習の手法を提案している[7)–10)]．藤田らは，トランプを 52 枚用いるため巨大な状態空間となること，相手の所持するカードは観測できないため部分観測状況となること，4 人対戦のマルチエージェントゲームであることの 3 つを，Hearts における戦略学習の困難性と考察している．そのうえで，困難性の解決手法として，パーティクルフィルタ，相手の行動予測器，状態を評価する状態価値関数，ゲームの特徴に基づく次元圧縮を提案している．計算機実験として，提案手法に基づく学習エージェントと，ルールベースエージェント 3 体とを対戦させた結果，約 2,000 ゲーム学習後にはルールベースエージェントよりも強い戦略を，約 4,000 ゲーム学習後には，人間の熟達者よりも優れた戦略を得ることに成功している．藤田らは，提案手法による戦略学習が Hearts だけではなく，他の部分観測カードゲームへも応用が可能であり，問題に依存した様々な状況に対しても適用することができると検討している．

2.3 従来研究の学習対象と戦略型ビデオ TCG の違い

戦略型ビデオ TCG の戦略学習においても，ゲームのある状態を評価するための評価関数は必須である．また，相手の所持するカードが観測できない部分観測空間となり，状態空間が巨大となるため，パーティクルフィルタなどのサンプリング手法が必要不可欠である．そのため，人間よりも優れた戦略を得ている藤田らのアプローチは，本研究においても非常に有望である．

しかし，戦略型ビデオ TCG に欠かせない要素である，「カードの組合せ」，「魔法」，「罠」を扱うためには，将棋やハーツとは異なった学習フレームワークが必要である．以下に，戦略型ビデオ TCG に欠かせない 3 つの要素の具体例をあげる．

カードを準備する段階での「カードの組合せ」　戦略型ビデオ TCG には，モンスターが設定されたモンスターカード，特殊な効果や技が設定された魔法カード，自分や相手が特定の条件を満たすことで発動する罠カードが用意されている．モンスターカードには，そのモンスターの強さ（体力，攻撃力，防御力など）と同時に，モンスターどうしの相性が設定されている場合が多く，プレイヤは必然的にモンスターの組合せを意識する必要がある．つまり，カードに複数のパラメータが存在する，カードどうしの相性が重要である，魔法や罠を含めたカードの組合せに重点を置く必要がある，という点で，従来研究の学習対象と戦略型ビデオ TCG は異なる．

「魔法」などの特殊効果　「魔法」には様々な特殊効果が設定されている．たとえば，自分のモンスターの攻撃力が上がる，相手を徐々に弱らせる，相手を行動不能にするなどで

ある．これらは，長期にわたり効果が持続したり，モンスターのパラメータに依存しなかったりする場合が多い．従来研究の学習対象には，魔法などの特殊効果は存在していないため，従来研究と同様の評価関数を用いることはできない．

「罠」のようなある特定の条件を満たすことで発動する効果　TCG の「罠」には，様々な発動条件や効果が設定されている．たとえば，相手が魔法カードを使ったときにその効果を反射する，相手の攻撃を無効化する，自分のモンスターを復活させるなどである．これらは，自分や相手がある条件を満たしたときのみ発動するため，発動条件を満たす確率を推定することで，罠を設置するタイミングを決定しなくてはならない．戦略型 TCG 以外で，罠のような効果が豊富に存在するゲームは少ない．そのため，従来研究の評価関数では，罠の発動条件を満たす確率や，罠を設置するタイミングを決定するのは難しい．

3. 問題の設定

3.1 ゲームのルール設定

本稿で用いる戦略型カードゲームは，プレイヤ対 COM の対戦型ゲームとし，ポケモンや遊☆戯☆王のルールを基にして以下のように設定する（各カードには 1 体のモンスターが設定されている）．ポケモンの戦略的要素である，属性の相性，特殊な技，モンスターの入れ替え，また，遊☆戯☆王の戦略的要素である，魔法カード，罠カードなどを扱えるようにルールを決定している．

(1)　プレイヤ，COM は 15 体のモンスターの中から，3 体選択する．
(2)　選択した 3 体から，戦闘状態とするモンスター 1 体を選択する．残り 2 体を待機状態とする（互いに，戦闘状態のモンスターしか観測できない）．
(3)　戦闘状態のモンスターに対して，攻撃，特殊攻撃，状態異常攻撃，罠設置，入れ替えを指示する．
(4)　プレイヤ，COM の戦闘状態のモンスターが，相手の戦闘状態のモンスターに対して，指示された行動を行う．入れ替えの場合は，現在の戦闘状態のモンスターを，待機状態のモンスター 1 体と入れ替える．
(5)　攻撃され，モンスターの体力が 0 になれば，そのモンスターは戦闘不能とする．
(6)　どちらかのモンスターが 3 体とも戦闘不能になるまで，(3)，(4)，(5) を繰り返す（(3)，(4)，(5) をまとめて 1 ターンと呼ぶ）．

各モンスターには，体力，攻撃力，防御力，特殊攻撃力，特殊防御力，素早さの 6 つのパラメータが存在する．また，攻撃，特殊攻撃，状態異常攻撃のダメージは，ポケモンと同様の計算式を用いている．

表 1　モンスターの属性表
Table 1　Elemental affinity.

攻撃する側＼攻撃される側	火	水	雷	地	ノ
火	△	△	ー	○	ー
水	○	△	ー	○	ー
雷	ー	○	△	×	ー
地	○	ー	○	ー	ー
ノ	ー	ー	ー	ー	ー

ダメージ倍率
○：２倍
ー：１倍
△：1/2倍
×：０倍

戦略的な要素を付け加えるため，各モンスターには属性を設定する（**表 1**，ノはノーマル属性を表す）．モンスターの特殊攻撃のみに，属性によるダメージ倍率が反映される．属性の相性は，相互補完的に設定されているため，現在戦闘状態の相手モンスターに対して有利に戦えるモンスターを，手持ちのモンスター 3 体の中から選択する必要がある．

「魔法」などの特殊効果の一例として毒攻撃（状態異常攻撃）を設定する．毒攻撃は 1 ゲーム中で 1 回のみ使えるとし，毒攻撃をされたモンスターは毒状態となる．毒状態のモンスターは，自分が行動する際に最大体力の 1/8 のダメージを受ける．つまり，毒攻撃をするタイミングと HP を考慮して，毒攻撃の対象とする相手モンスターを決定する必要性が生じる．

「罠」のようなある特定の条件で発動する効果の一例として魔法反射を設定する．魔法反射の罠設置は 1 ゲーム中で 1 回のみ使えるとし，罠の有効期間は設置してから 3 ターン（自分の順番が 3 回まわってくるまで）とした．魔法反射の罠が有効な間に，相手モンスターが魔法（状態異常攻撃）を使った場合，自分のモンスターが受けるダメージを無効化し，すべて相手モンスターに反射する．つまり，罠の有効期間の間に，相手モンスターが魔法（状態異常攻撃）を使うかどうか推測したうえで，罠を設置するかどうか決定する必要性が生じる．

本稿で用いる戦略型カードゲームは実際の戦略型ビデオ TCG と比べ，状態空間の大きさに関してははるかに小さい．ポケモンでは，モンスター 400 体以上の中から 6 体選択，属性は 15 種類以上，魔法などの特殊効果は 10 種類以上，モンスターの隠しパラメータも複数存在，クリティカル攻撃（相手に 2 倍のダメージ）や攻撃のミスなどのランダム要素，など状態空間はかなり大きくなる．遊☆戯☆王では，モンスターカード，魔法カード，罠カー

ドをすべて合わせると 7,000 種類程度存在する．しかし，戦略型ビデオ TCG の戦略的要素に着目すると，自分のターンにとるべき行動の選択，カードを準備する段階でのカードの組合せの決定,「魔法」などの特殊効果,「罠」のようなある特定の条件で発動する効果など，戦略型ビデオ TCG に欠かせない要素の大半は実現できているといえる．

3.2　戦略学習法の検討

3.1 節で述べた戦略型カードゲームを，戦略学習対象の観点から整理すると以下のようになる．

巨大な状態空間を持つ　ゲームにおける 1 局面を特定するものを状態と定義する．状態は，プレイヤと COM の持っているモンスターの組合せだけでも約 20 億状態になるが，それに加えて，両者の戦闘状態のモンスター，体力が減ったモンスター，戦闘不能であるモンスター，魔法や罠の効果も考慮しなくてはならない．

部分観測空間となる　相手が所持するモンスターは，戦闘状態となるまで観測することができない．つまり，相手の所持するモンスターを推定する必要があり，状態空間はさらに巨大になる．

ゲームにおける価値の分配　各ターンで自分のモンスターが相手モンスターに攻撃をする際に得られる価値のほかに，一時的には不利だが次のターンで有利になる場合や，相手モンスターを倒した場合などの，ゲーム全体を通して見たときの将来的な価値（遅れ価値）が重要な役割を占める．そのため，遅れ価値を過去にさかのぼって分配する必要がある．

戦略を学習するうえで有効な学習法はいくつか提案されており，研究例も少なくない．たとえば，隠れマルコフモデル（HMM）[11]，リカレント型ニューラルネットワーク（RNN）[12]，ベイジアンネットワーク[13] などがあげられる．しかし，巨大な状態空間による計算量の増大，遅れ価値が重要であることを考慮に入れると，本稿の戦略型カードゲームにおける学習法として適しているとはいえない．そこで，本研究では強化学習法[14] を用いる．モンスターの属性相性から予測されたダメージ量を，各ターンにおける推定即時報酬とする．戦闘履歴の即時報酬からゲーム状態における状態価値を算出し，学習を進めることで，遅れ価値を十分に反映させることができる．部分観測に起因する巨大な状態空間の問題は，相手の手持ちモンスターを推定したうえで，考慮すべきゲーム状態の候補の中からランダムにサンプリングすることで解決できる．ゲームの状態には，ゲームの特徴を考慮した状態圧縮を施すことで無駄な情報を省き，効率良く学習を進めることができるようにする．また，TCG 特有の要素である “最適なカード組合せ” や “魔法や罠などの特殊効果” に関しても，戦闘履歴のデータを用いることで学習することができる．強化学習法は，ゲーム中に得られる報酬から，試行錯誤を通じて戦略を学習するため，人間のエキスパートより優れた戦略を得られる可能性が期待されている学習法である[7)–10)]．

本稿の戦略学習機構は，上記の困難性を解決したうえで，最適行動の選択，最適組合せの選択，状態異常攻撃，罠設置の学習を目指す．

4.　学習機構の実装

4.1　最適行動選択

最適行動学習機構は，各ターンでの最適行動選択を目的とし，以下の 6 部分から構成される（**図 1**）．

1. 各ターンでの最適行動を決定する効用関数

効用関数 $U(H_t, a_t)$ は，ゲーム中のある局面（H_t はゲーム開始から，あるターン数 t までの戦闘履歴）におけるコンピュータの “行動 a_t” の価値を出力する．COM がとりうるすべての行動（攻撃，特殊攻撃など）において効用関数を求め，効用関数が最大となる行動を，ある局面における最適行動 $\pi(H_t)$ として出力する．つまり，最適行動が，効用関数の出力を最大化するように学習を進めることになる．

$$\pi(H_t) = \arg\max_{a_t} U(H_t, a_t) \tag{1}$$

効用関数は，以下の式により求める．ここで，s_t とは t ターン目の状態（局面），S はゲームのルール上とりうる状態の候補，$R(s_t, a_t, s_{t+1})$ は行動 a_t によって現状態 s_t から次状態 s_{t+1} に遷移する際に得る即時報酬，$V(s_{t+1})$ は次状態の状態価値を示す．

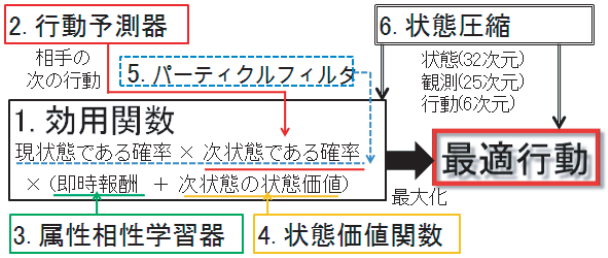

図 1　最適行動学習機構
Fig. 1　Learning method for optimum action.

$$U(H_t, a_t) = \sum_{s_t \in S} P(s_t|H_t) \sum_{s_{t+1} \in S} P(s_{t+1}|s_t, a_t)\{R(s_t, a_t, s_{t+1}) + V(s_{t+1})\} \quad (2)$$

しかし，本稿の戦略型カードゲームは部分観測ゲームであるため，COM は “状態” のすべて（プレイヤの持っているモンスターなど）を特定できない．そこで，“状態” のうち COM が知ることのできる情報を，COM の “観測 o_t” として定義する．効用関数は，プレイヤのモンスターを推測することで，COM の “現観測” から真の “現状態” を推定する．現状態の各候補には “現状態である確率 $P(s_t|H_t)$”（0.0〜1.0 の値をとる）が存在するため，その総和を $\sum_{s_t \in S}$ で求めている．しかし，“現状態である確率” の分布を正確に求めるためには，実際の戦闘履歴から統計的に算出する必要がある．そのため，現段階では，“現状態である確率” は一様分布であると仮定し，“1/現状態の候補数” として計算している．また，“行動” とはモンスターに指示する “攻撃” や “入れ替え” などのことを指す．“状態”，“観測”，“行動” に関しては「6. ゲームの特徴による次元圧縮」で詳しく述べる．

2. 相手の行動を予測する行動予測器

“現状態から次状態に遷移する確率 $P(s_{t+1}|s_t, a_t)$” は，相手の行動を予測する行動予測器により推定する．行動予測器は，効用関数によって推定された現状態を入力として，相手の行動 $\acute{a}_t$ の選択確率 $P(\acute{a}_t|s_t)$ を出力する（攻撃する確率，特殊攻撃する確率など）．次状態とは，現状態にプレイヤの行動と COM の行動を適用し，ゲームのルールに則して遷移した状態，と定義する．つまり，プレイヤの行動と COM の行動が決定すれば，次状態は一意に決まる．そのため，行動予測器により求めたプレイヤの行動選択確率が “現状態から次状態に遷移する確率（0.0〜1.0）” と一致する．

$$P(s_{t+1}|s_t, a_t) = P(\acute{a}_t|s_t) \quad (3)$$

しかし，本稿の戦略型カードゲームは部分観測ゲームであるため，コンピュータが相手の行動を予測する際は，$P(s_{t+1}|s_t, a_t) \approx \sum_{o_t \in O} P(\acute{a}_t|o_t)P(o_t|s_t)$ となる．

行動予測器の学習には，ニューラルネットの 1 つである多層パーセプトロン（MLP）を用いる．本研究で用いた MLP は，入力層，中間層，出力層からなる 3 層ニューラルネットであり，入力層は現状態の真の状態（32 次元），中間層は入力層と同じ 32 次元，教師は実際にプレイヤがとった行動（9 次元）として学習を進める．ニューロンの閾値関数には，0.0〜1.0 を出力するシグモイド関数を，また，ニューラルネットワークの訓練には，教師あり学習技術であるバックプロパゲーション（誤差逆伝搬学習法）を用いる．

3. 属性相性の学習器

属性相性学習器は，コンピュータの行動と，行動予測器によって予測した相手の行動との，属性相性倍率（0.0〜2.0）を出力する．たとえば，火属性が水属性に攻撃すると，2.0 倍のダメージを与えることができる（表 1）ため，属性相性学習器の出力は 2.0 に近い値を出力する．“現状態から次状態に遷移する際に得る即時報酬 $R(s_t, a_t, s_{t+1})$” は，相手に与えることができる推定ダメージと，相手から受ける推定ダメージの差とする．属性相性学習器が出力する属性相性倍率は，相手に与えるダメージと，相手から受けるダメージの推定に必要不可欠である．属性の相性は，本稿の戦略型カードゲームにおいて最も重要な要素であり，次状態に遷移した際の属性の相性は，ゲームの勝敗を大きく左右する鍵となる．属性の学習にも MLP を用い，入力層はプレイヤと COM のモンスターの属性（5+5 次元），中間層は入力層と同じ 10 次元，教師は通常のダメージと比べ何倍のダメージを与えられたか（属性相性倍率と等しくなる，1 次元），として学習する．

4. 状態の価値を推定する状態価値関数

状態価値関数は，“次状態の状態価値 $V(s_{t+1})$” を出力する．効用関数によって推定された現状態から，行動予測器によって推定された次状態に遷移した際，その次状態の戦況は有利なのか，不利なのかを求めている．具体的には，ゲームの状態を入力とし，その状態からゲーム終了までに得られる推定即時報酬の和を出力する．状態価値関数の出力が大きいゲーム状態ということは，今後得られるであろう即時報酬が多い，つまり，戦況は有利ということになる．状態価値関数の学習にも MLP を用い，1 ゲームが終わるごとにその履歴から学習する．履歴を用いることで，ある状態の価値を，各ゲームの終了時から逆向きに分配し学習することが可能となり，ゲームにおける価値に遅れを十分に考慮できる．入力層は現状態（32 次元），中間層は入力層と同じ 32 次元，教師はその時刻からゲーム終了までの推定即時報酬の和（1 次元），として学習を進める．

5. ランダムサンプリングによる状態推定

本稿で扱う戦略型カードゲームは，巨大な状態空間を持ち，かつ，部分観測ゲームとなる．そのため，現状態の候補 $s_t \in S$ と，そこから遷移する次状態の候補 $s_{t+1} \in S$ が大量に存在し，そのすべてにおいて効用関数を計算するのは不可能である．そこで，現状態の候補と次状態の候補から，いくつかの候補をランダムにサンプリングし，近似現状態 $\widehat{s_t}$ と近似次状態 $\widehat{s_{t+1}}$ を推定する．現状態，次状態の確率分布に基づいてサンプリングをするためには，実際の戦闘履歴を用いて統計的に算出する必要がある．現段階では，現状態，次状態の確率分布は一様分布であると仮定し，ゲームのルールに則している状態の中からランダムにサンプルを抽出していることになる．抽出された状態のみを用いて効用関数を計算し得られた結果は，すべての次状態から得られるはずの結果を近似的に表していることになる．

$$U(H_t, a_t) \approx \sum_{\widehat{s_t} \in S} P(\widehat{s_t}|H_t) \sum_{\widehat{s_{t+1}} \in S} P(\widehat{s_{t+1}}|s_t, a_t)\{R(\widehat{s_t}, a_t, \widehat{s_{t+1}}) + V(\widehat{s_{t+1}})\} \qquad (4)$$

6. ゲームの特徴による次元圧縮

戦略型カードゲームのゲーム状態の次元は，15 体のモンスターの組合せに加えて，モンスターの属性，各モンスターの体力や攻撃力などのパラメータ，モンスターが戦闘不能かどうか，などすべてを考慮すると非常に高い次元となる．そこで，上記 1〜5 における，“状態”，“観測”，“行動” にはゲームの特徴による次元圧縮が施されている．次元圧縮によって無駄な情報を省くことで，戦略学習機構は効率良く学習することが可能となる．以下に “状態”，“観測”，“行動” の圧縮方法を示す．

状態（32 次元）

0〜17 次元は，戦闘状態モンスターに関する情報．

0–4：COM のモンスターが，火，水，雷，地，ノのどれか？

5–9：プレイヤのモンスターが，火，水，雷，地，ノのどれか？

10，11：COM，プレイヤの体力

12，13：COM，プレイヤの素早さ

14：COM の攻撃力－プレイヤの防御力

15：プレイヤの攻撃力－ COM の防御力

16：COM の特殊攻撃力－プレイヤの特殊防御力

17：プレイヤの特殊攻撃力－ COM の特殊防御力

18–22：COM が所持する火，水，雷，地，ノ属性のモンスター数

23–27：プレイヤが所持する火，水，雷，地，ノ属性のモンスター数

28，29：COM，プレイヤの死んでいるモンスター数

30，31：COM，プレイヤのモンスターの体力の合計

0–4 次元，5–9 次元はそれぞれ，火，水，雷，地，ノに対応している．COM の戦闘状態モンスターの属性が火属性の場合，0 次元目は 1.0，1–3 次元目は 0.0 となる．また，18–22 次元，23–27 次元も同様に，火，水，雷，地，ノに対応している．COM が火属性モンスター 3 体を持っている場合，18 次元目は 3.0，19–22 次元目は 0.0 となる．

観測（25 次元）

“状態” の 32 次元から以下を取り除いたもの．

18–22：COM が所持する火，水，雷，地，ノ属性の数

30，31：COM，プレイヤのモンスターの体力の合計

行動（9 次元）

0：攻撃

1：特殊攻撃

2：状態異常攻撃

3：罠設置

4：火属性モンスターがいれば入れ替え

5：水属性モンスターがいれば入れ替え

6：雷属性モンスターがいれば入れ替え

7：地属性モンスターがいれば入れ替え

8：ノ属性モンスターがいれば入れ替え

0–8 次元ともに，0.0 または 1.0 の値であり，ブール変数として扱う．たとえば，COM が攻撃をする場合，0 次元目は 1.0，1–7 次元目は 0.0 となる．

“状態”，“観測” の次元圧縮では，重要な特徴に対して多くの次元を割り当てることで，より効率の良い学習が可能である．本稿の戦略型カードゲームの場合は，戦闘状態モンスターの属性と，所持しているモンスターの属性がきわめて重要であるため，COM とプレイヤにおいてそれぞれ 5 次元ずつ割り当てている．また，“状態” の 10–17 次元目と 30–31 次元目，“観測” の 10–17 次元目の値に関しては対数をとっている．他の次元がブール変数や一桁の実数であるのに対し，2 桁から 4 桁の実数となるため，MLP の学習が効率良く進まない可能性があるからである．

4.2　最適組合せ選択

最適組合せ学習機構は，ゲーム開始時におけるモンスター 3 体の最適な組合せの選択を目的とし，以下の 3 部分から構成される．

1. モンスター特徴推定器

モンスターの最適な組合せを選択する場合，属性の相性に加えて，モンスターの特徴の相性も考慮しなくてはならない．そこで，モンスターの特徴を 6 つ用意し，戦闘履歴から得たモンスターの戦闘データを用いてそれぞれ算出する（速攻攻撃型，速攻特殊攻撃型，重火力攻撃型，重火力特殊攻撃型，攻撃防御型，特殊攻撃防御型）．たとえば，速攻攻撃型は “攻撃で与えたダメージの平均 × 先に攻撃した確率”，重火力攻撃型は “攻撃で与えたダメージの平均 × 戦闘不能までに受けた攻撃の回数”，攻撃防御型は “攻撃されて受けたダメージの平均 × 戦闘不能までに受けた攻撃の回数”，として算出する．つまり，攻撃力と素早さが高い

モンスターは速攻攻撃型となることを意味する．モンスター特徴推定器にも MLP を用い，入力層はモンスターの固有の 6 つのパラメータ（体力，攻撃力，特殊攻撃力，防御力，特殊防御力，素早さ．6 次元），中間層は入力層と同じ 6 次元，教師は戦闘履歴から算出したモンスターの特徴の値（速攻攻撃型などの 6 次元），として学習を進める．つまり，そのモンスターの速攻攻撃型らしさ，特殊攻撃防御型らしさ，などを出力することになる．

2. モンスター組合せ特徴の次元圧縮

組合せ特徴（デッキタイプ）を考える際，モンスターの全組合せを考慮すると非常に高次元になってしまう．そこで，デッキタイプには次元圧縮を施し，10 次元まで圧縮する（火属性，水属性，雷属性，地属性，ノ属性，速攻型，重火力型，攻撃型，特殊攻撃型，守備型）．火属性デッキからノ属性デッキまでは，該当するモンスターの数である．速攻型デッキから守備型デッキまでは，モンスター特徴推定器で求めたモンスター特徴の値の和とする．

3. モンスター組合せの有効値推定器

モンスター組合せの有効値推定器も MLP を用い，1 ゲームが終わるごとにその戦闘履歴から学習する．入力層は，COM のモンスター 3 体の組合せとプレイヤのデッキタイプ（15+10 次元），中間層は，入力層と同じ 25 次元，教師は，COM がそのゲームに勝利した場合は 1.0 を，敗北した場合は 0.0 を与える（1 次元）．つまり，COM のモンスター 3 体の組合せにおいて，プレイヤのデッキタイプと対戦したときの勝利確率を出力するように学習を進めることになる．たとえば，COM のモンスター 3 体の組合せは，火属性デッキに対して勝利する確率は 80%だが，守備型デッキに対して勝利する確率は 20%である，という出力をする．モンスター 3 体の全組合せにおける有効値推定器の出力の中から，勝利確率が最大となるものを選択すれば，プレイヤのデッキタイプに対するモンスター 3 体の最適組合せを決定することができる．これは，モンスター 3 体の情報を有効値推定器に直接与えるのではなく，デッキタイプの次元圧縮を施した値のみを与えるだけで，最適組合せを決定できることを意味する．

市販の戦略型ビデオ TCG においては，相手のモンスター 3 体が何かは分からないが，相手のデッキタイプ（相手モンスター 3 体の組合せの特徴）は大体分かっていることが多々ある．そのため，相手のデッキタイプをモンスター組合せの有効値推定器に与えることは妥当である．仮に，相手のデッキタイプすら分からない場合は，属性相性の推定値と，モンスター特徴の相性の推定値のみを用いて，モンスター 3 体の最も相性の良い組合せを用いる．

4.3　状態異常攻撃の学習

TCG の「魔法」には様々な特殊効果（状態異常）がある．たとえば，毒，麻痺，睡眠などである．これらは，長期にわたり効果が持続したり，モンスターのパラメータに依存しなかったりする場合が多い．つまり，本稿における状態異常攻撃は，攻撃や特殊攻撃とは異なる特徴を持つことになる．そのため，状態異常攻撃に関する学習機構を新たに実装する必要がある．

状態異常攻撃の学習機構も MLP を用い，入力層は状態異常攻撃の対象となるモンスターの 6 つのパラメータと現在の残り体力（6+1 次元），中間層は入力層と同じ 7 次元，出力層はそのモンスターを毒状態にしたときのダメージの増加量（1 次元），となるように学習を進める．つまり，出力が大きくなるモンスターに対して状態異常攻撃をしたほうが，状態異常の効果を最大限発揮させることができる，ということである．状態異常の効果は長期にわたり効果が持続するため，対象となるモンスターの現在の残り体力を入力として与え，状態異常攻撃をするタイミングを学習している．最適行動学習機構（4.1 節）において最適行動を決定する際に，状態異常攻撃の学習機構の出力が最大となる場合は状態異常攻撃を選択する．これにより，状態異常攻撃をするタイミングと，状態異常攻撃の対象とする相手モンスターを決定できる．学習機構には状態異常の効果に関していっさい教えておらず，ダメージの増加量のみを用いて有効値を学習させているため，毒以外の様々な状態異常に関しても同様に学習できる．

4.4　罠設置の学習

TCG の「罠」には，様々な発動条件や効果が設定されている．たとえば，魔法反射，攻撃の無効化，モンスターの復活などである．これらは，自分や相手がある条件を満たしたときのみ発動するため，発動条件を満たす確率を求めることで，罠を設置するタイミングを決定しなくてはならない．つまり，本稿における罠は，攻撃や特殊攻撃，また，状態異常攻撃とも異なる特徴を持つことになる．そのため，罠に関する学習機構を新たに実装する必要がある．

罠設置の学習機構にも MLP を用い，現状態を入力とし，相手モンスターが未来 3 ターンの間に状態異常攻撃をする確率を出力するように学習を進める．つまり，出力が大きいときに罠を設置すれば，相手が罠にはまる可能性は高いことになる．相手モンスターが未来 3 ターンの間に状態異常攻撃をする確率は，ゲーム終了時の戦闘履歴を用いて，入力層はゲームの現状態（32 次元），中間層は入力層と同じ 32 次元，教師は，相手が未来 3 ターンで状態異常攻撃をすれば 1.0，しなければ 0.0 として学習する（1 次元）．罠の効果自体は魔法と似たものが多いため，状態異常攻撃の学習機構を用いて学習する（魔法反射の罠が発動した場合，相手に毒攻撃をすることになる）．最適行動学習機構（4.1 節）において最適行動を

決定する際には，罠の効果が相手に有効かどうかを状態異常攻撃の学習機構の出力から決定し，そのうえで，罠設置の学習機構の出力が最大となる場合に罠設置を選択する．これにより，罠の有効期間内に発動条件を満たすかどうかを推測したうえで，罠の効果が相手に有効であるならば罠を設置する，という決定が可能になる．

5. 計算機実験

本稿における戦略学習機構に基づく学習エージェント（RL-agent）を，ルールベースエージェント（Rule-based）と対戦させることで，戦略学習機構の有効性を評価する．

RL-agent が Rule-based 相手に 100 ゲーム学習するごとに，同じ Rule-based と 200 ゲームの評価ゲームを行い，RL-agent の勝率などを求める．本稿の戦略型カードゲームの勝敗は，モンスター 3 体の組合せに大きく依存するため，Rule-based が評価ゲームで用いるモンスターの組合せは，あらかじめ 200 ゲーム分をランダムに決定しておく．RL-agent のモンスターの組合せは最適組合せ学習機構が決定する．一方，学習ゲームにおけるモンスターの組合せは，RL-agent，Rule-based ともに，完全にランダムで決定した．

本稿における戦略学習機構が，プレイヤの様々な戦略に適応可能かどうか調べるため，Rule-based の個性として，攻守のバランスがとれた “バランス型”，攻撃志向の “力押し型”，防御志向の “堅実型” を用意した（以降，バランス型 Rule-based，力押し型 Rule-based，堅実型 Rule-based）．RL-agent と，3 種類の Rule-based との実験結果は 5.2 節で述べる．

次に，戦略学習機構が，新たなルールの追加に臨機応変に対応できるかどうかを調べるため，最初は状態異常攻撃と罠設置を禁止し，途中から使用可能にした場合において実験する．新たなルールの追加に関する実験結果は 5.3 節 で述べる．

最後に，市販のビデオ TCG に，本稿の戦略学習機構を応用するための準備段階として，本稿の戦略型カードゲームにおける汎用的な戦略に関して検証する．汎用的な戦略に関する実験結果は 5.4 節で述べる．

5.1 Rule-based のルール

実験に用いたバランス型 Rule-based は以下の 11 個のルールを持ち，(1) を最優先ルールとして，以下順番に優先度が低くなるよう設定されている．

(1)　まだ罠を使っておらず，かつ，現状態で罠を使うことが有効ならば，罠設置．
(2)　まだ毒攻撃をしておらず，かつ，相手モンスターに対して毒攻撃が有効ならば，状態異常攻撃．
(3)　自分の戦闘状態モンスターの属性（自属性）が，相手の戦闘状態モンスターの属性（敵属性）に対し「○」ならば特殊攻撃．
(4)　自属性が敵属性に対し「○」となるモンスターに入れ替え．
　　敵属性が自属性に対し「○」ならば，
(5)　「×」となるモンスターに入れ替え．
(6)　「△」となるモンスターに入れ替え．
(7)　「−」となるモンスターに入れ替え．
　　自属性が敵属性に対し「×」ならば，
(8)　「−」となるモンスターに入れ替え．
(9)　「△」となるモンスターに入れ替え．
(10)　自属性が敵属性に対し「△」ならば，「−」となるモンスターに入れ替え．
(11)　攻撃か特殊攻撃か，ダメージの大きい方を選択．

個性を変えたものとして，攻撃志向の “力押し型”Rule-based はルール 8〜10 をルール 5 よりも優先度が高くなるように設定してあり，できる限り相手に多くのダメージを与えられるような行動を選択する．また，“堅実型”Rule-based はルール 5〜7 をルール 4 よりも優先度が高くなるように設定してあり，できる限り自分がダメージを受けないような行動を選択する．

5.2 様々な戦略への適応性

戦略学習機構に基づく学習エージェント（RL-agent）を，バランス型 Rule-based，力押し型 Rule-based，堅実型 Rule-based 相手に，5,200 ゲーム学習させた際の RL-agent の勝率を**図 2** に示す．グラフは，RL-agent の 3 回の学習過程における，RL-agent の勝率の 500 ゲーム間の移動平均であり，横軸は学習したゲーム数，縦軸は RL-agent の勝率を表す（以降，グラフはすべて同じ条件で表す）．

計算機実験の結果，学習をしていない段階での RL-agent の勝率は 25 約 2,200 ゲーム学習後には 80%程度まで上昇した．よって，戦略学習機構は正常に戦略を学習していることが確認できた．RL-agent の勝率が 50%を超えるのは約 500 ゲーム学習後であり，戦略学習機構は比較的早い段階で，Rule-based と拮抗する程度の戦略を得ることができている．また，力押し型 Rule-based，堅実型 Rule-based との対戦においても，バランス型 Rule-based とほぼ同様の結果が得られた．どの Rule-based と対戦した場合も，RL-agent の勝率が 80%程度まで上昇したことから，Rule-based の個性が変化しても，対戦相手の戦略に応じて戦略

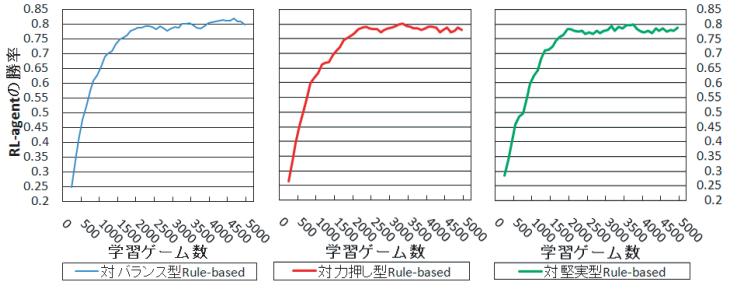

図 2　様々な Rule-based との対戦結果
Fig. 2　RL-agent's winning percentage.

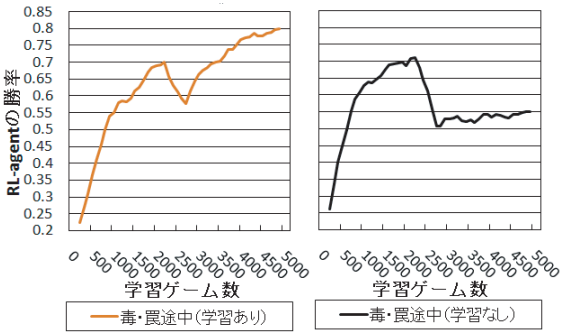

図 3　途中で毒と罠を追加
Fig. 3　Addition of new rules.

学習ができているといえる．

5.3　新たなルールの追加への適応性

前節の実験結果から分かるように，RL-agent の勝率は，2,200 ゲーム学習後に 80%付近で収束している．そこで，RL-agent，Rule-based ともに，最初は状態異常攻撃と罠設置を禁止しておき，2,500 ゲーム学習後から使用を許可した場合の計算機実験を実施した．

図 3 左は状態異常攻撃と罠に関して学習をする場合，図 3 右は状態異常攻撃と罠に関して学習をまったくしない場合である．状態異常攻撃と罠を戦略学習機構に学習させた場合（図 3 左），RL-agent の勝率は 2,500 ゲームを機に 1 度下がるが，すぐに上昇に転じ，最終的には図 2 と同じ 80%程度となる．しかし，状態異常攻撃と罠を戦略学習機構に学習させなかった場合（図 3 右），RL-agent の勝率は上昇しない．よって，途中で新しいルールを追加した場合でも，戦略学習機構はルールの変化に適応できていることが示された．また，図 3 左において，状態異常攻撃と罠設置を許可する前では RL-agent の勝率は 70%程度だが，許可することで勝率は 80%程度まで上昇するようになる．つまり，ゲームのルールが複雑になることで，より戦略性が増していることが分かる．

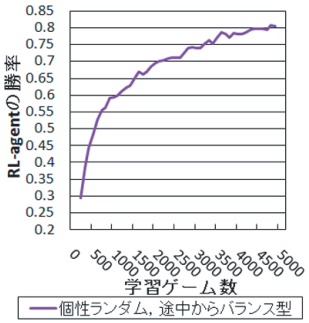

図 4　個性ランダム（途中からバランス型）
Fig. 4　Select RL-agent at random.

5.4　汎用的な戦略の獲得

本稿の戦略型カードゲームにおける汎用的な戦略を得るために，学習ゲーム時，評価ゲーム時の Rule-based の個性を，3 種類（バランス型，力押し型，堅実型）の中からランダムに決定する Rule-based との対戦を実施した．2,500 ゲームまでは，Rule-based の個性をランダムに決定し，2,500 ゲーム学習後は，Rule-based の個性をバランス型で固定した場合の実験結果を**図 4** に示す．

2,500 ゲームの時点では RL-agent の勝率は 70%程度にとどまり，バランス型 Rule-based と対戦したときの 2,500 ゲーム時点での勝率（図 2 左）より 10%程度低くなった．しかし，2,500 ゲーム以降は徐々に勝率が上昇し，最終的に RL-agent の勝率は図 2 左と同様に 80%程度となる．この結果から，2,500 ゲーム目までは，戦略型カードゲームにおける，相手の戦略に依存しない部分の戦略を獲得していることになる．つまり，戦略型カードゲームの汎用的な戦略を得ている．2,500 ゲーム学習後は，汎用的な戦略を基にして，相手の戦略へと徐々に適応できている．

6. 検　　討

6.1 戦略学習機構の検討

計算機実験の結果から，人間プレイヤと対戦する TCG において，本稿で提案する戦略学習機構を利用できることが示された．バランス型，力押し型，堅実型など，人間プレイヤの様々な戦略に応じた戦略を獲得できる．人間プレイヤが急に戦略を変更したり，新たな「魔法」や「罠」を使用したとしても，戦略学習機構は新たなルールの追加に対して適応できる．また，比較的早期に相手の戦略と拮抗する戦略を得られていることと，最終的には 80%の勝率が得られていることから，プレイヤの要求に合わせた COM の強さの設定が可能である．

現段階における戦略学習機構の課題としては，現状態の推定と，ランダムサンプリングによる状態推定の際に，現状態である確率，次状態である確率の確率分布を一様分布と仮定している点があげられる．そのため，最適行動選択における学習が収束しにくい可能性がある．現状態である確率と，次状態である確率を得るためには，人手で経験的知識から状態遷移の確率を決定するか，あるいは，良い戦略を持ったプレイヤ同士の戦闘履歴から，統計的に現状態の尤度と次状態の尤度を求める必要がある．

最適行動，最適組合せ，魔法の効果，罠の効果を自動的に学習できることにより，戦略型ビデオ TCG に欠かせない要素の大半を実現できているといえる．しかし，モンスターの使用にある条件を満たさなければならない場合の自動学習は実現できていない．たとえば，マジック：ザ・ギャザリングには土地（ランド）カードが存在し，モンスターごとに定められた枚数以上の土地カードを持っていなければ，そのモンスターを召喚できない．また，遊☆戯☆王などの生贄（アドバンス）召喚や特殊召喚，儀式召喚など，モンスターの召喚に他のモンスターを生贄にしたり，魔法カードからモンスターを召喚したりする場合もある．モンスターの使用に条件があるゲームの戦略学習を実現するためには，まず，どのモンスターを使用したいのか決定する機構が必要である．さらに，そのモンスターを召喚する条件を満たすための戦略学習が必要不可欠である．そして，最終的に得られた遅れ価値を，「モンスターを召喚する条件を満たす」部分に分配しなくてはならない．

6.2 市販ビデオ TCG への実応用

本稿で提案する戦略学習機構は，相手の戦略から自らの学習パラメータを自動的に決定し，相手の戦略よりも良い戦略を獲得することができる．また，TCG の汎用的な戦略を得たうえで，それを基にして相手戦略に順応させていくことも可能である．

実際の市販ビデオ TCG に応用する場合，人間プレイヤの戦略に対してリアルタイムに追従する必要がある．リアルタイム追従を実現する 1 つの可能性として，5.4 節の実験結果は非常に有効である．あらかじめ TCG 全体の汎用的な戦略を獲得しておき，それを基に相手戦略へと順応させることで，相手の戦略に応じた強い COM を短期間で実現できる．もう 1 つの可能性としては，まず，数パターンのルールベース戦略に対して，戦略学習機構を用いてあらかじめ学習パラメータを決定しておき，次に，数ゲームでプレイヤの戦略がどのパターンの戦略か判別し，それに応じた学習パラメータに切り替える．数ゲームの間は，相手の戦略に対して戦略学習機構で学習を進めたうえで，あらかじめ決定しておいた学習パラメータの中から，一番近いものを判別手法を用いて決定すれば解決できる．急激な戦略の変化を避けたいならば，あらかじめ決定しておいた学習パラメータと，現在の学習パラメータの中間値をとればよい．あとは，戦略学習機構により，相手の戦略から学習パラメータを随時学習させれば，リアルタイム性を疑似的に実現できる．

戦略学習機構によって相手の戦略よりも良い戦略を獲得できたとしても，市販ビデオ TCG においては，人間プレイヤの要求に合わせて COM の戦略を作る必要がある．たとえば，人間プレイヤが初心者の場合は，当然弱い COM の戦略が要求される．また，中級者以上でも，弱い COM と戦ってストレスを発散する，あえて敵を弱くし邪魔させないようにして自分の腕を磨く，などのために COM の戦略を細かく要求する可能性もある．つまり，戦略学習機構によって獲得した戦略を，わざと弱くして利用しなければならない．そのためには，どの学習パラメータがどの程度勝率に影響するのかを調べたうえで，人間プレイヤの要求に合わせて，学習パラメータを調整する必要がある．

7. おわりに

本稿では，戦略型ビデオ TCG を題材とした戦略の学習機構を提案し，計算機実験により評価と検討を行った．本稿の戦略型ビデオ TCG は，巨大，かつ，高次元の状態空間を持つが，パーティクルフィルタと，次元圧縮により，計算量を削減した．また，行動予測器，属性相性推定器，状態価値関数により，真の状態を推定することで，最適行動を選択する機構を設計した．加えて，モンスター特徴推定器，モンスター組合せの有効値学習器により，最適なモンスター組合せを学習する機構を設計した．TCG における「魔法」などの特殊効果の例としては，状態異常攻撃を設定した．状態異常攻撃を行った際の有効値を学習させることで，状態異常攻撃のタイミングと，毒攻撃の対象とする相手モンスターを決定する機構を設計した．TCG における「罠」のようなある特定の条件で発動する効果の例としては，魔法反射を設定した．罠の有効期間内に相手モンスターが魔法（状態異常攻撃）を使うかどう

か推定したうえで，罠を設置した際の優位性を学習する機構を設計した．また，計算機実験により，戦略学習機構が正常に動作していること，比較的早期に相手の戦略と拮抗する戦略を得られること，新たなルールの追加に対して適応できていること，戦略型カードゲームの汎用的な戦略を獲得できることが確認された．戦略型ビデオ TCG において，プレイヤ同士の駆け引きを学習するための足がかりとして，相手の戦略に臨機応変に対応する戦略学習機構を実現できた．

本研究により，ルールやパラメータによって COM の強さを作り込むのではなく，対戦によって戦略を学習し変化させる，汎用的な COM の実現可能性が示された．また，従来研究ではなされていない，カードの組合せ，魔法や罠などの特殊効果といった，ビデオ TCG 特有の要素においても，自動的に戦略を学習できるフレームワークを示すことができた．ゲームの多様性，ひいては，面白さを確保するための手段の 1 つとして本研究の応用が期待される．

今後は，より効率良く学習を進めるための手法を検討することで，プレイヤの戦略対してより短期間で適応する戦略学習機構を構築する必要がある．また，Web 上でプレイすることが可能なゲームシステムを構築し，人間プレイヤと戦略学習エージェントとを対戦させることで，戦略学習機構の有用性を示す必要がある．

参　考　文　献

1) 遊☆戯☆王，コナミ (1999). http://www.yugioh-card.com/japan/
2) Guinness World Records. http://www.guinnessworldrecords.com/
3) ポケットモンスター，ポケモンカードゲーム，任天堂 (1996). http://www.pokemon.co.jp/
4) 玉越勢治，八木昭宏，片寄晴弘ほか：fNIRS を用いた対戦型ゲームのエンタテインメント性の初期的検討—対人間と対コンピュータにおける比較，情報処理学会研究報告，Vol.2006, No.134, 2006-EC-3, pp.29–35 (2006).
5) 塩入健太，星野准一：仮想対戦プレイヤーの感情的発話生成，インタラクション 2006 論文集，pp.157–164 (2006).
6) 保木邦仁：コンピュータ将棋における全幅探索と futility pruning の応用，情報処理学会学会誌，Vol.47, No.8, pp.884–889 (2006).
7) 藤田　肇，石井　信：マルチエージェントカードゲームのための強化学習法の改良，電子情報通信学会技術研究報告，Vol.102, No.731, pp.167–172 (2003).
8) Ishii, S. and Fujita, H.: A Reinforcement Learning Scheme for a Partially-Observable Multi-Agent Game, *Machine Learning*, Vol.59, pp.31–54 (2005).
9) 藤田　肇，石井　信：部分観測カードゲームのためのモデル同定型強化学習，電子情報通信学会論文誌，Vol.J88-D-II, No.11, pp.2277–2287 (2005).
10) Fujita, H. and Ishii, S.: Model-Based Reinforcement Learning for Partially Observable Games with Sampling-Based State Estimation, *Neural Computation*, Vol.19, pp.3051–3087 (2007).
11) 高野　渉，山根　克，中村仁彦：運動の認識・生成に基づく原始的コミュニケーションの階層構造モデル，日本ロボット学会学術講演会予稿集 (2005).
12) 尾形哲也，小嶋秀樹，駒谷和範，奥野　博：RNNPB による視聴覚情報変換を利用したロボットの身体・音声表現，電子情報通信学会技術研究報告，TL-2006-22, NLC-2006-18, PRMU2006-99, pp.45–50 (2006).
13) 湯浅将英，安村禎明，新田克己：オンライン交渉における擬人化エージェントの表情選択支援，ヒューマンインタフェース研究会報告，Vol.2004, No.74, 2004-HI-109, pp.1–6 (2004).
14) Sutton, R.S. and Barto, A.G.: *Reinforcement Learning, An Introduction*, MIT Press (1998).
15) 北川源四郎：モンテカルロ・フィルタおよび平滑化について，統計数理，Vol.44, No.1, pp.31–48 (1996).

(平成 21 年 3 月 19 日受付)
(平成 21 年 9 月 11 日採録)

藤井　叙人（学生会員）

2007 年関西学院大学理工学部卒業．2009 年関西学院大学大学院理工学研究科修士前期課程修了．株式会社野村総合研究所に就職．当時の研究テーマ「テレビゲームにおける戦略学習」．

片寄　晴弘（正会員）

1991 年大阪大学大学院基礎工学研究科博士課程修了．工学博士．イメージ情報科学研究所，和歌山大学を経て，現在，関西学院大学理工学部教授．ヒューマンメディア研究センターセンター長．音楽情報処理，感性情報処理，HCI の研究に従事．科学技術振興機構さきがけ研究 21「協調と制御」領域研究者．科学技術振興機構 CREST「デジタルメディア（略称）」領域 CrestMuse プロジェクト研究代表者．電子情報通信学会，人工知能学会各会員．