様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　５　月　　８　日現在

機関番号：　３２６１９
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３０４１９

研究課題名（和文）方策勾配法によるマルコフ決定過程を前提としない強化学習の理論とゲームへの応用

研究課題名（英文）Theoretical research of the policy gradient reinforcement learning without Markov properties and its application to games

研究代表者

五十嵐　治一（Harukazu, Igarashi）

芝浦工業大学・工学部・教授

研究者番号：８０２８８８８６

交付決定額（研究期間全体）：（直接経費）　　1,500,000円

研究成果の概要（和文）：本研究は強化学習の一方式である方策勾配法において、方策関数の表現法と学習方式を考案し、ゲーム分野や工学的応用への方法論を構築することを目的として理論的な研究と応用面での研究を行った。この結果、次の研究成果を得た：
(1)方策勾配法の理論的な研究として、①階層化による高度な戦略の学習方式の提案、②環境ダイナミクスと行動知識の分離学習の提案、③方策としてファジィ制御ルールを用いた場合の学習方式の提案をすることができた。
(2)方策勾配法の応用面の研究として、追跡ゲーム、ロボットサッカー、コンピュータ将棋等への適用を試み、提案手法の有効性を確認することができた。

研究成果の概要（英文）：In this research project, we have made theoretical and practical research for developing expressions of policy functions and learning methods in the policy gradient reinforcement learning algorithms. Our final goal is constructing a general methodology that can be applied to computer games and engineering fields. The results of this project are as follows.
(1)Theoretical research on the policy gradient reinforcement learning: we proposed new methods in ① hierarchical reinforcement learning to learn higher strategies of agents, ②learning with separated knowledge of environmental dynamics and action-values in agent policies, and ③learning with a fuzzy controller for policies.
(2) Practical application of the policy gradient reinforcement learning: we applied the proposed learning methods to pursuit games, robot soccer games and computer shogi and examines the efficiency of our methods.

研究分野：　人工知能

キーワード：　強化学習　方策勾配法　マルチエージェント　コンピュータ将棋　ロボカップ　ソフトマックス探索

１．研究開始当初の背景

従来､強化学習はマルコフ決定過程（MDP）への適用が殆どであり、非マルコフ決定過程への適用はあまり進んでいない。その中で、方策勾配法は MDP の仮定を必要としない強化学習法として注目されている。研究代表者らも方策勾配法が、状態遷移、報酬、方策に関する３つのマルコフ性の仮定が成立しない一般的な非マルコフ決定過程へと拡張できることを示してきた。

特に、方策の表現として、IF-THEN ルールの重み（行動価値に相当）や､状態の価値、ポテンシャル等のヒューリスティクスを用いることや、それらの合成が可能であることを示した。さらに、IF-THEN ルールとしてファジィ表現を可能とする方法も提案してきた（ファジィ推論と方策勾配法の融合方式）。このように、Q 学習などの他の強化学習の諸手法と比べて、方策勾配法は方策の表現として多様な形式を用いることが可能な点が大きな強みである。

２．研究の目的

本研究プロジェクトは、方策勾配法における方策関数の表現法と学習方式を考案し、ゲーム分野や工学的応用への方法論を構築することを目的としている。これを実現するために次の研究を行うことを計画した：

(1)方策の階層化の理論的な研究：
状態空間､行動空間､方策の階層化を考え、学習の高速化を図るとともに、戦略レベルなどの高度な状況判断や行動決定の学習を可能にする方式を考案する。

(2)環境ダイナミクスと行動知識の分離学習：
環境ダイナミクス（エージェントの動作特性）と対象とする問題を解決するための行動に関する知識とを方策中で分離して表現し、両者ともに学習する学習方式を実際に事例に適用してその有効性を検証する。

(3)ファジィ推論と方策勾配法との融合方式：
前件部と後件部にファジィ表現を含み、ルール重みを持ったファジィ推論システムにおいて、方策勾配法により前件部と後件部のメンバシップ関数とルール重みを学習させる方式を考案する。また、入力／出力変数が離散的である場合と連続的である場合の両方にも対応する。

(4)ゲームに関するヒューリスティクスを不要とするゲームエージェントの学習方式：
将棋などの探索を要するゲームにおいて、方策勾配法により対局のみから局面評価関数を学習させる方式を考案する。特に、学習の高速化のための近似法も開発する。

３．研究の方法

前章の研究目的(1)～(4)を達成するために以下の方法をとる：

(1)階層化と方策表現についての従来法の調査と新方式を考案する。

(2)学習実験による理論の検証。応用例題として、非マルコフ的な走行特性を持つエージェント群による追跡問題を対象とする。

(3)サッカーロボット（RoboCup 小型リーグ）の行動決定問題を例題とした学習実験と、自動車の速度制御の学習問題をシミュレーション上で行う。

(4)コンピュータ将棋への方策勾配法の適用を試みる。学習方式の考案、学習の高速化のための近似手法の考案や、将棋での局面評価関数とシミュレーション方策の学習実験を行う。

４．研究成果

２章､３章で述べた研究項目(1)～(4)についての研究成果を順に述べる。

(1)方策の階層化の理論的な研究
すでに２階層からなる行動決定モデルについて、動作モデルを設計し、方策勾配法の学習則の導出はできあがっていた。本研究プロジェクトでは、これを一般的な n 階層のモデルへ拡張し、その学習則の導出の詳細を論文にまとめて発表した〔雑誌論文 1〕。

(2)環境ダイナミクスと行動知識の分離学習
マルチエージェント系の強化学習に対する代表的なベンチマーク問題の一つである追跡問題でグリッドの サイズが 7×7 でハンター数が 2 の問題を対象とした。標準的な環境ダイナミクスの下で問題解決に必要とされる状態行動対の価値（行動価値）から成る行動知識パラメータと、対象世界の環境ダイナミクスを推定した状態遷移パラメータとによって構成される目的関数の具体的な関数系を提案し、方策勾配法を用いた両パラメータの学習則を導出した。提案方式を追跡問題の事例へ適用する実験を行った結果、適切な行動選択が可能な方策を従来法と同じく学習することができた。また、行動知識と環境ダイナミクスとを分離して取り扱い、既知の環境ダイナミクスを利用することで、学習に要するエピソード数を従来法より少なくすることが可能であることを確認した。さらに、分離した行動知識の転移学習により、環境ダイナミクスが変化した場合の学習を従来法より効率化できるという実験結果を得た〔雑誌論文 2〕。

(3)ファジィ推論と方策勾配法との融合方式
ロボットサッカーの世界的競技会である Robo Cup Japan Open 2012（2012 年 5 月, 大阪）における小型リーグの試合の中から、ロボットとボールの配置を再現した 30 シー

ン（静止画像）について、そこでの行動決定問題を学習対象とした。図 1 に一例を示す。図 1 は“ シーン 16 ”の例で、○印で囲ったロボットはボールを保持しているロボットとその味方ロボットで、 0－2 でリードされている局面である。ここで直接シュートを打つか、味方へパスを出すか、あるいは、ドリブルによりボールを運ぶかの 3 種類の行動の中から一つを選択したい。このような状況判断はコンピュータよりも人間の方が優れていると考えられる。しかし、人間の判断を具体的なアルゴリズムや厳密な数値で予め表現しておくのは難しい。それよりもいくつかの特徴量で試合局面を表現し、ファジィ表現を含んだルールによる行動決定方式が向いている。また、人によって判断結果も異なることがあり、そのばらつきの分布を反映させた報酬を考えて強化学習を行った。学習実験の結果、30 シーン中、25 シーンまではほぼ完全に学習することができた〔雑誌論文 5〕〔学会発表 7〕。

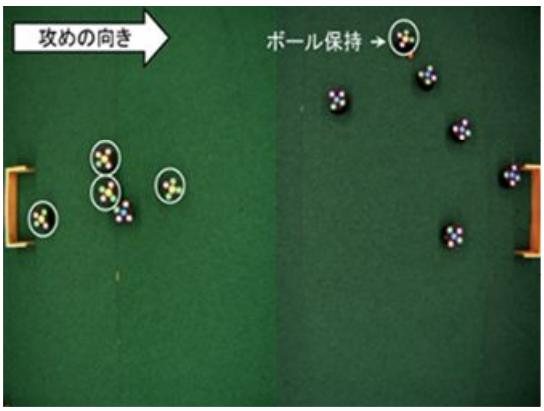

図 1　学習対象とした 30 シーンの一例：シーン 16. ○印で囲ったロボットはボールを保持しているロボットの味方のロボットである。

また、我々は従来から RoboCup サッカーシミュレーションリーグ 2D 部門においても。パッサーやレシーバの行動決定に方策勾配法を適用してきた。前述の小型リーグにおける研究成果にヒントを得て、これと同様に人間が試合観戦中にプレイや局面を評価して、報酬を学習エージェントへ与えることによりオンラインでプレイヤの局面評価関数を学習するシステムを考案・開発した(図 2)。

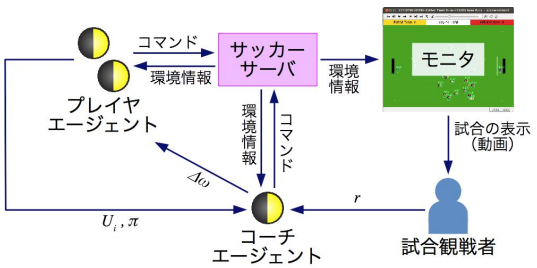

図 2　サッカーエージェントのオンライン学習システムの構成

本システムを用いて学習実験を行った結果、学習前のチームと比べてスルーパスの出現回数や成功回数が増加することや、敵ゴール前でのパス回しからシュートする回数が増え、勝率や平均得点が大幅に増加する可能性があることが確認できた〔学会発表 1,2,4,9,10〕

(4)ゲームに関するヒューリスティクスを不要とするゲームエージェントの学習方式

コンピュータ将棋において、プロ棋士の棋譜データベースを用いることなく、自己対戦による勝敗情報だけから局面評価関数を学習する研究を行った。まずは、従来から行われてきた強化学習のチェスやコンピュータ将棋への適用例について調査した。チェスにおいては TD(λ)法や TDLeaf (λ)法が成果を上げているが、コンピュータ将棋では未だに大きな成果を挙げるまでには至っていないことがわかった。

そこで、コンピュータ将棋において、プロ棋士の棋譜データベースを全く用いないで、コンピュータが自己または他者との対局のみを通じて局面評価関数を学習し、棋力向上を図る方法について考察した。その結果、学習エージェント自身との自己対局や他者との対局を行い、勝敗や主観的評価、探索における最善応手手順や自分の対局譜から、強化学習や教師付学習を用いて局面評価関数を学習し、棋力を向上させるなどの方法を提案した。

特に、この中で、①マルコフ性を仮定することなくエピソード単位の VAPS アルゴリズムを適用すること、②VAPS アルゴリズムにおける目的関数をエピソード収益とエピソード誤差の線形和とすることを提案し、③エピソード収益を勝敗情報のようにマルコフ性のある報酬（ステップ報酬）と、戦法や陣形、駒の選好などの棋風に基づく主観的評価を表現したマルコフ性のない部分報酬（エピソード報酬）とに分離して与えた場合の学習則を導出し、④これらの枠組みで TD(λ)法や方策勾配法の学習則が導出でき、これまでに将棋に適用されてきた強化学習の殆どの学習則が導出できることを示し、さらに、⑤同じ枠組みで教師付学習法の学習則を導出し、この際に VAPS アルゴリズムの生成項も考慮することを提案した〔雑誌論文 3〕〔学会発表 5,6,8〕。

さらに、ソフトマックス戦略に基づくシンプルな探索方式を提案し、コンピュータ将棋へ適用した実験を行った。本探索方式では探索木中のノードの評価値は子ノードの評価値を選択確率で重み付けした期待値であり、再帰的に定義される。選択確率は選択先のノードの評価値を目的関数とするボルツマン分布を用いる。探索は実現確率を良さの度合いとする最良優先探索であり、深さの制御には実現確率の閾値を用いた反復深化を用いる。各ノードへの実現確率はルートノードからの選択確率の積で定義する。したがって、将棋の有効な指し手に関するヒューリスティクスは使用せず、最終的には局面評価関数だけに依

存する。従来方式のプログラムと対戦させた結果、約 43%の勝率を得た。従来プログラムの Min-Max 戦略の中で行われている枝刈りなどの複雑な高速化手法を全く取り入れていない割には高い勝率を得た〔学会発表 3〕。

５．主な発表論文等

〔雑誌論文〕（計 5 件）

1. Harukazu Igarashi and Seiji Ishihara, "Hierarchical Policy Gradient Reinforcement Learning: Two-layer Model," The Research Reports of Shibaura Institute of Technology, Natural Sciences and Engineering, 査読無, vol.60, no.2, pp.21-28 (March 31, 2017), ISSN 0386-3115. DOI:10.13140/ RG.2.2.19842.89285

2. 石原聖司, 五十嵐治一：" 方策に関する知識を分離した方策こう配法―環境ダイナミクスと行動価値による方策表現― "，電気学会論文誌 C（電子・情報・システム部門誌），査読有，Vol.136, No.3, pp.282-289 (2016) , DOI:10.1541/ieejeiss.136.282.

3. Harukazu Igarashi, Yuichi Morioka and Kazumasa Yamamoto, "Learning Positional Evaluation Functions without Using Databases of Game Records between Professional Shogi Players," The Research Reports of Shibaura Institute of Technology, Natural Sciences and Engineering, 査読無, vol.59, no.2, pp.39-47 (March 31, 2016), ISSN 0386-3115. DOI: 10.13140/RG.2.1. 4797.2242

4. Harukazu Igarashi, Yuichi Morioka and Kazumasa Yamamoto, "Reinforcement Learning of Positional Evaluation Functions and Simulation Policies by Policy Gradient Algorithm," The Research Reports of Shibaura Institute of Technology, Natural Sciences and Engineering, 査読無, vol.58, no.1, pp.1-9 (March 31, 2015), ISSN 0386-3115.

5. 杉本将也, 五十嵐治一, 石原聖司, 田中一基：" ファジイ制御ルールにより表現された方策を持つ方策勾配法：RoboCup 小型リーグにおける行動決定 "，知能と情報（日本知能情報ファジィ学会誌），査読有, Vol. 26, No. 3, pp.647-657 (June,2014), DOI:10.3156/jsoft. 26. 647

〔学会発表〕（計 10 件）

1. 田川 諒，五十嵐治一，" サッカーエージェントにおけるスルーパスの強化学習 "，第 15 回情報科学技術フォーラム講演論文集 (FIT2016)， 2016 年 9 月 7 日，富山大学(富山県富山市)

2. 大内 斉，五十嵐 治一，" 局面評価関数を用いたサッカーエージェントの移動先決定 "，第 21 回ゲーム・プログラミング・ワークショップ, 2016 年 11 月 4 日，箱根セミナーハウス（神奈川県足柄下郡箱根町）

3. 原悠一, 五十嵐治一, 森岡祐一, 山本一将，" ソフトマックス戦略と実現確率による深さ制御を用いたシンプルなゲーム木探索方式 "，第 21 回ゲーム・プログラミング・ワークショップ, 2016 年 11 月 4 日，箱根セミナーハウス（神奈川県足柄下郡箱根町）

4. 田川諒，五十嵐治一，" サッカーエージェントにおける局面評価関数の強化学習 "，第 20 回ゲーム・プログラミング・ワークショップ, 2015 年 11 月 6 日，箱根セミナーハウス（神奈川県足柄下郡箱根町）

5. 大串明, 山本一将, 森岡祐一, 五十嵐治一，" コンピュータ将棋における方策勾配を用いた局面評価関数の教師付学習 "，第 20 回ゲーム・プログラミング・ワークショップ, 2015 年 11 月 6 日，箱根セミナーハウス（神奈川県足柄下郡箱根町）

6. 五十嵐治一，森岡祐一，山本一将，"プロ棋士の棋譜データベースを用いない局面評価関数の学習法についての考察 "， 第 34 回ゲーム情報学研究発表会, 2015 年 7 月 4 日，九州工業大学サテライト福岡天神（福岡市中央区）

7. Noor Imanina N.H. , and Harukazu Igarashi, "Policy Gradient Method Using Fuzzy Controller in Policies and Its Application," Proceedings of the International Conference on Artificial Intelligence and Pattern Recognition (AIPR2014), Nov. 17-19, Kuala Lumpur (Malaysia)

8. 五十嵐治一，森岡祐一，山本一将，" 方策勾配法による探索制御の一考察 "，第 19 回ゲーム・プログラミング・ワークショップ 2014, 2014 年 11 月 7 月, 箱根セミナーハウス（神奈川県足柄下郡箱根町）

9. 田川 諒, 谷川俊策, 五十嵐治一，" agent2d のチェーンアクションにおける評価関数の重み調整 "，第 13 回情報科学技術フォーラム講演論文集(FIT2014), 2014 年 9 月 3 日，筑波大学（茨城県つくば市）

10. 谷川俊策，五十嵐治一，石原聖司，" RoboCup サッカーシミュレーションリーグ 2D における局面評価関数の設計と学習 "，ロボティクス・メカトロニクス講演会 2014,

2014年5月26日, 富山市総合体育館(富山県富山市)

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
五十嵐　治一（IGARASHI, Harukazu）
芝浦工業大学・工学部・教授
研究者番号：80288886

(2)連携研究者
石原　聖司（ISHIHARA, Seiji）
東京電機大学・理工学部・准教授
研究者番号：50351656

(3)研究協力者
森岡　祐一（MORIOKA, Yuichi）
山本　一将（YAMAMOTO, Kazumasa）