RoboCup2D プレイヤのファジィ方策関数の進化的獲得
Fuzzy Rule Genetic evolution for RoboCup simulation 2D player

西野順二、佐土瀬寛
Junji NISHINO、 Kan SADOSE
電気通信大学
The University of Electro-Communications
nishinojunji@uec.ac.jp

Abstract

サッカープレイヤの行動決定手法としてファジィ方策関数を提案し、その方策パラメータであるファジィ集合を GA を用いて進化的に獲得した。獲得実験では、世代集団、親個体群、子個体群の個体数をそれぞれ 810、10、20 と設定した。ランダムな初期方策から始めて agent2d との対戦結果を適応度とし 1200 世代以上の進化実験を行った結果、世代ごとの適応度である累計得失点の平均値が約-2.5 点から +2.0 点超まで増加し、強いチームを得ることができた。この計算を実行するため、40 ワーカからなるサッカーシミュレーションクラスタを構築し用いた。獲得したファジィ方策関数を実装したプログラムと agent2d による対戦実験の結果 、 学習前のプログラムの勝率 23 %と比較して 54.9 %と高い勝率であったことから、GA による進化的獲得の有効性を明らかにした。

1 はじめに

強化学習の枠組みでは環境に対する行動戦略を方策関数 π で表し 、θ によるパラメータ化を行い $\pi(\theta)$ とおく。最適なパラメータ の発見を目指すことが行われている。例えばニューラルネットであれば θ はリンクの重み変数に相当し、重み付きのルールベースシステムならば、ルールの重みとルールの記述パラメータに相当する。このとき $\Delta\theta$ を明示的に計算することでを更新する学習測が用いられる。

しかしながら サッカープレイヤの協調行動のように、状態空間が大きく、報酬がスパースかつ遅延が大きいという複雑な問題の強化学習での解決は非常に困難である。

DQN[Mnih 13] は、状態空間をニューラルネットによって近似しながら学習を行うことで、巨大な状態空間への対応を行い簡単なゲームでは高い学習性能をあげている。しかし、報酬がスパースで遅延するものは学習がうまくできていないことと、学習ができた場合でも獲得されたニューラルネットの解釈・分析が事実上できないという課題がある。

ファジィ方策関数は、複雑な問題の大局的な解法知識をファジィ推論によって表現することで、こうした困難な問題を可読性のある形で解決することを目指している。ファジィ推論規則で表現した方策関数は ニューラルネットに比べて構造的な記述能力が高いという特徴がある。一方でファジィ推論の非線形性と非単調性（多峰性）、いたるところで微分不可能という特性から、$\Delta\theta$ を明に得ることができず、勾配を用いた最適化によって方策を求められない。

そこでヒューリスティックなメタ探索である遺伝的アルゴリズム [北野 93] を用いてファジィ方策関数を獲得することとした。

本研究では疎なサッカーシミュレーション用クラスタを構築し 40 試合を並列で行うことでより進化世代数を大幅に増やすことを試みた。近年、強化学習では試行をより多数、この場合 GA の進化世代数をより長期間行うことで、よりよい解を得られることが示されている [Silver 17]。遺伝的アルゴリズムによるサッカールールの獲得は、2006 年ころまで積極的に行われてきた。このころ Taylor らは 420 時間相当 (約 2,500 試合)[Taylor 06]、Luke らは 10,000 試合相当の学習 [Luke 98] を行っている。計算資源の制約によりこの程度の実験が行われていた。本実験では、大幅に長期間にあたる 8,000 時間相当 (48,000 試合）の進化反復を専用のサッカークラスタを構築して実施した。

本研究では RoboCup Soccer simulation 2D リーグを対象とし、サッカープレイヤのファジィ方策関数を提案しそのパラメータであるファジィ集合を GA を用いて学習しファジィ方策関数を獲得することを目的とする。

2 RoboCupSoccer2D プレイヤにおけるファジィ方策関数

2.1 RoboCupSoccer2D シミュレーション

本研究では、RoboCup Soccer simulation 2D リーグのプレイヤを対象とする。simulation 2D リーグのシミュレータは、サーバクライアント方式で構築され、rcssserver と呼ばれる一つのサッカーサーバプログラムと、両チーム 11 ずつのプレイヤに相当する、のべ 22 クライアントプログラムで 1 つの試合を実行する。試合時間は 6000 ステップ、実時間で 10 分である。

サッカーサーバでは 2 次元平面で高さのない仮想サッカーフィールドを用意し、ボールやプレイヤの物理的な運動のシミュレーションを行う。サーバは、各クライアントのサッカーエージェントプログラムへ、仮想の視覚・聴覚情報の送信、物理計算、状態更新など 100 ミリ秒単位で行う。サッカーエージェントはこの知覚情報等を元に状況を判断し、100 ミリ秒以内に行動を決定し、行動コマンドをサーバへ送信する。

2.2 ベースプログラム agent2d

本研究では ベースプログラムとして agent2d を用いた。これは秋山らによって開発された 2010 年の世界大会優勝チーム HELIOS の基本的な行動をライブラリ化したキットのサンプルエージェントプログラムである。

ドリブルなどの基本行動や、状態予測と探索に基づく行動計画、ボール位置を反映したチームフォーメーション、基本的なセットプレー、ペナルティキックなどのその他の基本行動がビルトインされている。優勝チーム HELIOS にごく近く、サンプルでありながら、そのままの状態で、標準的なサッカーチームとしては十分な強さを持っている。

2.3 chain action 探索

agent2d では ボール保持者の行動を決定する方策決定として chain action 探索を用いている。

探索では以下のアルゴリズムによって行動決定を行う。

1. 初期状態を入力とし行動の候補を生成する。
2. 成功と予測される行動を生成し 行動とその結果の予測局面を組としたノードを作る。
3. ノードを生成するときに予測局面を評価し 各ノードは評価値を保持する。
4. 最良優先探索により終了条件を満たすまで木を成長させる。
5. 探索木の生成が終了した後 評価値が最も高いノードにつながるアクション連鎖を決定する。

ここで生成される探索木を図に示す。

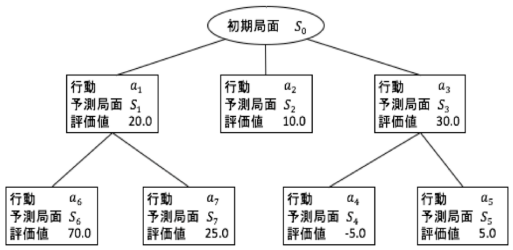

図 1: chain action の探索木

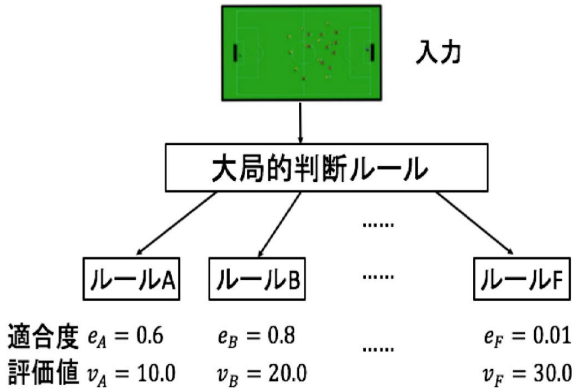

図 2: ファジィ局面評価関数

2.4 ファジィ方策関数

ファジィ方策関数 $\pi(s)$ は、状態から行動を決定するためのアルゴリズムとして、ファジィ推論を組み込んだ方策関数である。状態に対して行動を確定的に決定する。

谷川ら、田川らは chain action 探索内で用いる評価関数を人間の知識に基づく状態特徴量の線形和で表現し、これに対して強化学習を行った [谷川 13, 田川 15] 。

ファジィ方策関数は、人の直感的知識を推論構造に反映し、線形和に比べて表現力が高い特徴がある。

ファジィ方策関数は、谷川らと同様に、chain action 探索を基本形として採用し、探索で用いる評価関数を多段ファジィ推論で表現した。これは、ボール保持者の行動を決定するとき、大局的な状況判断とその状況に応じた状態評価の二段階に対応している。

ファジィ推論の構造を図 2 に示す。

3 ファジィ方策関数の進化的獲得

ファジィ方策関数は、ファジィ推論に依拠する非線形な複雑さにより、一般に多峰的であり、確率的方策勾配法など一般的な強化学習手法によってパラメータを学習することが困難である。本研究では ファジィ方策関数のファジィ推論パラメータを GA を用いて獲得することを提案した。GA の世代交代モデルには JGG を、交叉手法にはファジィ推論パラメータが実数値であることから、REX

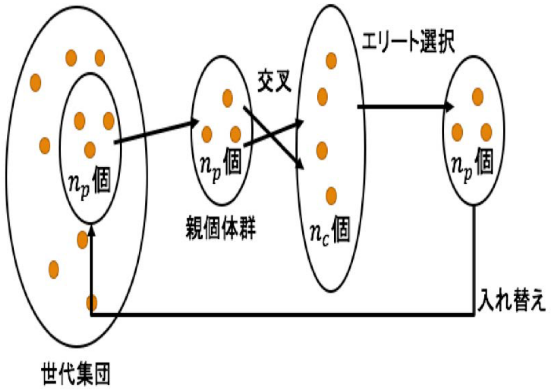

図 3: 世代交代モデル JGG (Just Generation Gap)

を採用した [小林 09]。

3.1 世代交代モデル JGG (Just Generation Gap)

JGG (Just Generation Gap) 法にもとづく世代交代の手順は以下のとおりである。

1. 与えられた初期化領域に n_{pop} 個の個体をランダムに生成し初期集団とする。
2. 集団からランダムに n_p 個の親個体を非復元抽出する。
3. 親個体集団に交叉を繰り返し適用し n_c 個の子個体を生成する。
4. 子個体の適応度を評価する。
5. 子個体集団を適応度にしたがってソートし、上位 n_p 個体を親個体と置き換えて次世代の集団とする。

この流れを図 3 に示す。

3.2 交叉 REX (Real-coded Ensemble Crossover)

交叉 REX では式 (2) および 式 (1) にしたがって子個体 $\mathbf{x}^c \in \mathcal{R}^n$ を生成する 。$\mathbf{p}^g$ を親個体群の重心ベクトルである。

$$\mathbf{p}^g = \frac{1}{n_{pop}} \sum_{i=1}^{n_{pop}} \mathbf{p}^i. \tag{1}$$

$$\mathbf{x}^c = \mathbf{p}^g + \sum_{i=1}^{n_{pop}} \xi_i (\mathbf{x}^i - \mathbf{p}^g),\ \xi_i \sim \phi \tag{2}$$

4 ファジィ方策関数の進化的獲得実験

世代の個体数などの設定を表 1、子個体の評価に関する設定を表 2、のとおりとしてファジィ方策関数の進化的獲得実験を行った。

表 1: 集団の設定

次元数	54
世代集団数 n_{pop}	810
親個体群の個体数 n_p	10
子個体群の個体数 n_c	20

表 2: 子個体評価の設定

適応度	累計得失点差
評価試合数	2 試合
対戦相手	agent2d

4.1 サッカー GA クラスタ

GA によって進化実験をするとき、評価を子個体数だけ行う必要がある。世代数を千回、子個体を 10 として、1 試合で評価をするとしても 10,000 回の試行が必要となる。一方 RoboCup Soccer simulation 2D の試合は、1 試合のシミュレーションを 10 分で行うため、10,000 回ではのべ約 70 日を必要とする。

このため、本研究ではクラスタを構築し、進化計算の高速化を行った。ここで、子個体評価のためのサッカーシミュレーションが違いに影響しない疎な関係にあるため、独立な計算機を 40 台並列に用いた。設定は子個体数を 20 個体、評価は 2 試合であるため 1 世代での試合数は 40 となる。

本クラスタを用いることで、1 世代分のサッカーシミュレーションを一度に計算することができ、世代をまとめる GA 操作には時間がかからないため、全体として効率はほぼ 40 倍であった。詳細には、個々の試合ごとにラグタイムがあるため、各世代単位で終了待ちの同期を取った。シミュレーション全体への影響はほとんどなかった。

4.2 GA による進化と最良個体評価

以上の実験設定で、1200 世代超まで進化実験を行った。合計試合数は 1200x20x2 = 48,000 試合である。

世代ごとの適応度平均の推移を図 4 に示す。個体集団平均で -2.5 点程度の得失点差だったが 1.0 点程度まで上昇し agent2d 相手に強くなったことがわかる。

いくつかの世代のクライアントを取り出し agent2d と対戦させたところ、学習前 (初期世代集団) の個体の勝率は 23.2 %、学習後の最良個体は 54.9 %、手作業によるファジィ方策関数を用いたクライアントは 51.2 %の勝率であった。

以上の結果から、強いファジィ方策関数を獲得できたと言える。

また、評価の変化はまだ上昇をつづけている傾向があり、収束したとは言えず、従来より増やした 1200 世代でも進化には十分とはいえないことがわかった。

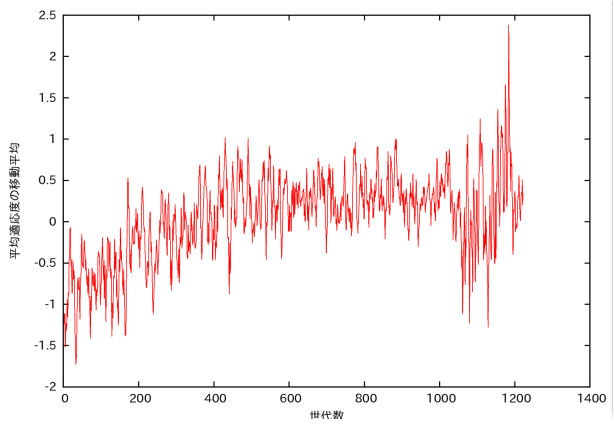

図 4: 進化過程 (1200 世代)　縦軸は 2 試合での累計得失点差

5　おわりに

本研究では、強化学習による行動学習が難しい問題に対し、人の直感を構造的に生かしたファジィ方策関数を設定し、そのパラメータを GA を用いて進化的に獲得する手法を提案し、実験した。

従来より多数の、1200 世代、48,000 試行以上の個体評価を専用クラスタを構築して行って進化実験をしたところ、適応度が世代を繰り返すごとに増加したことから GA によるファジィ方策関数の獲得の可能性を確かめることができた。

学習後のファジィ方策関数を用いたプログラムは基準の agent2d に対し勝率 54.9 %であり強いプログラムが得られている。

今後の課題として、まだ収束しきっていない進化をさらに続けることで、何世代まで進化させるのが適当であるかの知見を得ることが望まれる。

謝辞本研究の一部は科学研究費補助金基盤研究 C(26330273) および電気通信大学共同研究 (株式会社 QUANTUM) により支援されたものである

参考文献

[Luke 98] Luke, S., et al.: Genetic programming produced competitive soccer softbot teams for robocup97, *Genetic Programming*, Vol. 1998, pp. 214–222 (1998)

[Mnih 13] Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M.: Playing atari with deep reinforcement learning, *arXiv preprint arXiv:1312.5602* (2013)

[Silver 17] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al.: Mastering Chess and Shogi by Self-Play with a General Reinforcement Learning Algorithm, *arXiv preprint arXiv:1712.01815* (2017)

[Taylor 06] Taylor, M. E., Whiteson, S., and Stone, P.: Comparing evolutionary and temporal difference methods in a reinforcement learning domain, in *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pp. 1321–1328ACM (2006)

[小林 09] 小林重信：実数値 GA のフロンティア, 人工知能学会論文誌, Vol. 24, No. 1, pp. 147–162 (2009)

[谷川 13] 谷川俊策, 五十嵐治一, 石原聖司　他：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学習, ゲームプログラミングワークショップ 2013 論文集, pp. 106–109 (2013)

[田川 15] 田川諒, 五十嵐治一　他：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, ゲームプログラミングワークショップ 2015 論文集, Vol. 2015, pp. 78–83 (2015)

[北野 93] 北野宏明：遺伝的アルゴリズム, 産業図書 (1993)