7P-06

RoboCup サッカーシミュレーション 2D リーグにおけるニューラルネットワークを用いた評価関数の模倣学習

岡崎弘侑† 五十嵐治一†

芝浦工業大学†

1. 研究背景と目的

RoboCup サッカーシミュレーション２D リーグでは、ベースとなるチームとして「agent2d」[1]がよく用いられる。agent2d では、攻撃時のボール保持者の行動決定に「チェーンアクション」という方法が採用されている。チェーンアクションでは、行動を枝、行動後のフィールド状況をノードとする探索木を作成後、ノードを評価して最も評価の高いノードへつながる枝の行動を選択する。このときノードを評価する関数にはボール位置を用いている。

本研究では、ニューラルネットワークのようなヒューリスティクスを全く含まない評価関数を構築したい。しかし、学習パラメータが大量となることから、強化学習ではかなりの学習時間を要することが予想される。そこで、強いチームの行動決定を模倣することにより、短い時間での学習をめざした。

2. 関連研究

RoboCup サッカーのフィールド状況をニューラルネットワークで評価する研究として中島ら[2]の研究がある。そこでは、入力を試合中のフィールド画像、出力を次にゴールが決まるまでの時間をもとに計算される$SituationScore$の値に設定して、畳み込みニューラルネットワークを学習した。本研究では、入力としてボールとプレイヤの位置データを用いる。また、このように人間が定めた局面評価値を正解出力として与えることはなるべく避けたいと考えた

3. 本研究における模倣学習

状態sに応じて与える報酬$r(s)$と以下の式で表される状態価値関数$V^{\pi}(s)$を考える。

$$V^{\pi}(s) \equiv E_{\pi}\{\sum_{k=0}^{\infty} \gamma^{k}\, r(s_{t+k}) \big| s_t = s\} \qquad (1)$$

このとき次のベルマン方程式が成り立つ。

$$V^{\pi}(s_t) = E_{\pi}\{r(s_t) + \gamma V^{\pi}(s_{t+1})\} \qquad (2)$$

本研究では、強豪チームの方策π_0の下で(2)式を満足する状態価値関数$V^{\pi_0}(s_t)$の存在を仮定する。次に、$V^{\pi_0}(s)$をニューラルネットワーク$V(s;\omega)$で近似し、確率的勾配法により両者の２乗誤差の最小化を図った。ただし、$V^{\pi_0}(s_t)$の値は$r(s_t) + \gamma V(s_{t+1};\mathrm{w})$で置き換えた。

なお強豪チームの試合のログから、試合の各フレームにおけるボール、全プレイヤの位置座標、速度ベクトルを取得して状態 s とした。学習は試合中からエピソードを切り出し、エピソード単位の学習を行った。

4. 学習実験

4.1 学習データの作成

模倣対象として「HELIOS 2019」を選んだ。このチームと agent2d の試合を 500 試合行い、ログファイルから学習用のデータを作成した。

また、報酬rを学習対象チームがゴールを決めたときに+100 点を与えると定め、「エピソード」をゴール直前の 50 フレームとし、その結果、1908 個のエピソードを作成した。

4.2 ニューラルネットワークの構成

本研究のニューラルネットワークは、入力層として、ボールと学習対象のチームのプレイヤ 11 人、相手のチームのプレイヤ 11 人の位置座標、速度ベクトルを表す計 92 個のユニットを使用している。中間層は 1 層で 40 個のユニットから構成され、入力層と中間層は全結合されている。活性化関数としては ReLU 関数を使用した。

出力層は評価値$V(s;\omega)$を出力するユニット 1 個で構成されて、中間層の各ユニットと全結合されている。活性化関数は作用させていない。出力層が出力する値は次の式で表される。

$$\mathrm{V}(\mathrm{s};\omega) = \sum_{i}^{40} \omega_i^{(2)} h_i(x) + b^{(2)} \qquad (3)$$

h_iは中間層のi番目のユニットが出力する値、x_jは入力層のj番目のユニットからの入力を表す。

Imitation Learning of Evaluation Functions Using Neural Networks in RoboCup Soccer Simulation 2D Leagues

†Hiroyuki Okazaki, Harukazu Igarashi, Shibaura Institute of Technology

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

4.3 ネットワークの重みの更新

本研究では、以下の①と②の手順を繰り返すことで学習を進めた。（図 1）

①：エピソードに含まれる全てのフレームの状態sでの状態価値関数の目標値$V'(s_t)$を以下のように計算し記憶する。

$$V'(s_t) = r(s_t) + \gamma V(s_{t+1}) \tag{4}$$

このとき割引率γ =0.99 とした。

②：次の損失関数が最小になるように最急勾配法により重みωを更新する。

$$loss(s_t) = \left(V'(s_t) - V(s_t;\omega_t)\right)^2 \tag{5}$$

ただし、学習率の更新には Adam[3]を用いた。したがって、重みの更新式は以下のようになる。

$$\omega_{t+1} = \omega_t - \alpha * \frac{\widehat{m_t}}{\sqrt{\widehat{v_t}} + \epsilon} \tag{6}$$

ただし、$\widehat{m_t}$ と $\widehat{v_t}$ はそれぞれ $\nabla_\omega \mathrm{loss}(s_t)$、$(\nabla_\omega loss(s_t))^2$から計算される。パラメータはそれぞれ$\alpha = 0.001, \epsilon = 1.0 * 10^{-7}$とした。

この①、②の手順を 1908 個のエピソード全てを用いて行うことを 1 エポックと定め、本研究では 1500 エポックの学習を実行した。

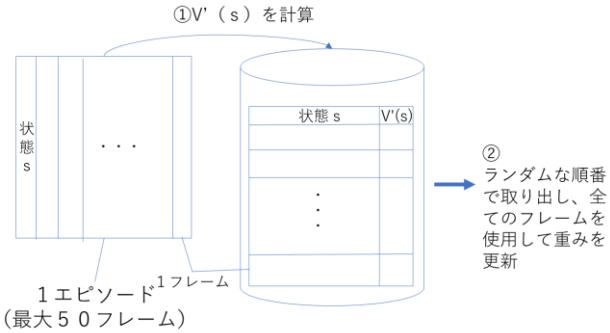

図 1　重みの更新手順

5. 実験結果と考察

本研究では、ニューラルネットワークの学習を同じ条件で 10 回行った。図 2 に学習時の損失関数と学習を行ったエポック数の関係を表す。

この図から、同じ 1500 エポックの学習を行った時点での損失関数の値で大きく 2 つのグループに分けることが出来る。ここで、損失関数の値が比較的大きいグループ（2、3、4、5、9、10 回目の実験結果）をグループ A、比較的小さいグループ（1、6、7、8 回目の実験結果）をグループ B とする。それぞれの学習結果をチェインアクション内で評価関数として使用するチームを作成し、それぞれ agent2d と 100 試合ずつ対戦を行った（表 1）。ただし、得点と失点は 1 試合あたりの平均を表し、勝率は引分けを除いて計算した。

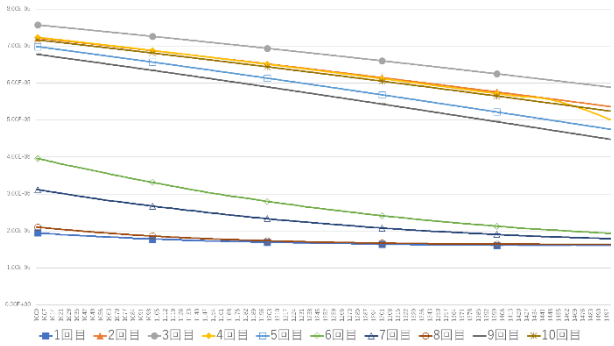

図 2 学習時の損失関数とエポック数の関係(1000 サイクル以降)

表 1　対 agent2d の対戦結果

	勝率	勝ち	負け	分け	得点	失点
A	8%	27	310	263	0.13	0.82
B	42%	127	176	95	1.74	2.09

表 1 からグループ B がグループ A より勝率が高く、平均得点が多いことが分かる。一方、グループ A は引き分けの数が多く、平均失点が少なかった。そこで、試合の内容を確認したところ、グループ A の試合では、相手のゴールの近くで一人のプレイヤがボールを一か所で保持し続けてしまい、試合が止まってしまっていることが分かった。これはある種の局所解に学習が陥ってしまっていると推測される。

6. まとめ

本研究では、強豪チームの試合ログから状態の遷移を取り出し模倣学習を行った。その結果、agent2d に対して約 42%の勝率となるような結果を得ることが出来た。しかし、局所解に近づいてしまい勝率が低くなることも確認された。

今後は、強豪チームの方策π_0と学習した状態価値関数を用いたチームの方策πを直接近づける学習を行うことで勝率を向上させるとともに、学習を安定させていきたい。

また、模倣学習で得られたチームをさらに強化学習で強化し、勝率の向上を目指すことも今後の目標として挙げられる。

参考文献

[1] 秋山英久, “RoboCup サッカー2D シミュレーションリーグ解説：サンプルエージェントを使ったチーム開発”, 知能と情報, vol.23, no.6, pp.838-844, 2011

[2] 中島智晴, POMAS Tanguy, “RoboCup サッカーシミュレーションにおける画像情報からのフィールド状態評価”, 第 34 回ファジィシステムシンポジウム, WB1-1, 2018

[3] Diederik P. Kingma, Jimmy Ba, “Adam: A Method for Stochastic Optimization”, https://arxiv.org/abs/1412.6980, 2015

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.