2S-03

RoboCup サッカーシュミレーション 2D リーグにおける 模倣学習の改良

岡崎弘侑† 五十嵐治一†
芝浦工業大学†

1. はじめに

RoboCupサッカーシュミレーション2Dリーグでのチーム開発のベースプログラムに「agent2d」[1]がある。agent2d では、攻撃時のボール保持者の行動決定の際に、「チェーンアクション」という仕組みを用いている。そこでは現在のフィールド状況から選択可能な行動を枝、行動後のフィールド状況をノードとする探索木を生成する。次に、各ノードの局面の優劣をスコア化し、最も評価の高いノードにつながる行動を選択する。

agent2d では評価関数としてボールの位置情報が用いられていたが、プレイヤの配置なども考慮するため、表現力の高いニューラルネットワークで近似することが考えられる。岡崎らは強豪チームの得点シーンのログから模倣学習によりこの評価関数を学習する研究を行った[2]。本研究ではこれに加えて、得点シーン以外のシーンについても教師あり学習により評価関数を改良することを試みた。

2. 先行研究

先の岡崎らの研究[2]では、評価関数をニューラルネットワークを用いて表現し、強豪チームの試合ログから模倣学習を行った。具体的には、得点シーンのログから状態の遷移データと設定した報酬を基に、TD(0)法と同様にベルマン方程式を利用した状態価値関数を計算した。

しかし、得られた状態価値関数をチェーンアクションの評価関数として組み込んでも勝率が向上せず、学習が局所解に陥った場合もあった。また学習チームの方策自体を強豪チームの方策に直接近づける操作までは行っていなかった。そこで、本研究では、この模倣学習で用いた試合ログから教師チームの行動を正解行動として取得し、学習チームが正解行動を選択するように模倣学習後に教師あり学習を追加した。

Improving Imitation Learning in RoboCup Soccer Simulation 2D Leagues
†Hiroyuki Okazaki, Harukazu Igarashi, Shibaura Institute of Technology

3. 教師あり学習

本研究では山岸らの教師あり学習の適用例[3]を参考にした。ただし、そこでは人間が試合の再生画面で正解行動を指示していた。本研究では試合ログから教師チームの行動を正解として用いる。以下、教師あり学習の概要を説明する。

まず、学習チームの行動選択の方策$\pi(a|s;\omega)$を(1)に示すような確率的方策とする。

$$\pi(a|s;\omega) \equiv \frac{e^{E(a,s;\omega)/T}}{\sum_{x\in A(s)} e^{E(x,s;\omega)/T}} \tag{1}$$

ここで、$E(a,s;\omega)$は局面sでの行動aを行った後の局面評価値である。

次に、(2)のカルバック・ライブラー情報量δ_{KLD}を最小化するように局面評価関数を学習する。ただし、$\pi^*(a|s)$は教師チームの方策である。

$$\delta_{KLD}(\sigma;\pi^*,\pi) \equiv \sum_{s\in\sigma}\sum_{a\in A(s)} \pi^*(a|s) ln \frac{\pi^*(a|s)}{\pi(a|s;\omega)} \tag{2}$$

このときの学習則は以下の式で表される[3]。

$$\Delta\omega = \frac{\varepsilon}{T}\sum_{s\in\sigma}\sum_{a\in A(s)} [\pi^*(a|s) - \pi(a|s;\omega)]\nabla_\omega E(a,s;\omega) \tag{3}$$

4. 学習実験

4.1 学習の流れ

今回の実験の流れは以下の通りである：

① 教師チームの試合ログを作成する。
② 作成したログから状況sとその時行った正解行動のペア(s,a^*)を取得する。
③ 状況sを再現し、チェーンアクションでの行動候補aの集合$A(s)$を生成する。
④ 学習則(3)によりωの更新を繰り返す。

ニューラルネットワークは文献[2]と同様で、入力層は状態sとして、ボールとプレイヤ(22 名)、の座標と速度ベクトルの計 92 個である。出力層は、その局面sの評価値を表す値が出力される。中間層は1層でユニット数は40、初期重みは文献[2]の模倣学習で得られた値にセットした。

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

4.2　試合ログの作成

今回の学習では、文献[2]の模倣学習で使用した試合ログをそのまま使用した。具体的には、模倣対象の「HELIOS 2019」と agent2d との 254 試合分の試合ログから、教師チーム(HELIOS)のプレイヤがボールを保持した状態sを 176490 個取得し、②～④で使用した。

4.3　正解行動a^*の取得

②のsにおける正解行動a^*をホールド、ドリブル、パス、シュートに判別する。例えば、kick を行ったプレイヤと次に kick を行ったプレイヤが同じ/異なるならドリブル/パスと解釈した。

4.4　学習プレイヤの行動候補$A(s)$の生成

③のsにおける学習プレイヤの行動候補$A(s)$はチェーンアクションにより生成する。しかし、必要な情報の全てがログに記録されているわけではない。学習プレイヤはログファイルからボールとプレイヤ(22 人)の座標と速度ベクトルをトレーナを介して取得するが、一定の待機時間中に得られた自己の聴覚・視覚情報と統合し、教師プレイヤが経験した局面sの再現を試みた。

次に、行動候補aの候補数を絞り込むため、同じプレイヤへのパスが複数個ある場合は、最も評価の高い候補のみを選択対象とした。その他、自分へのパスについてはその場にとどまるホールドと、移動するドリブル（パスと同じく一つを選択）とを行動候補とした。

4.5　教師チームの選択確率$\pi^*(a|s)$の計算

学習プレイヤの各行動候補aについては、次のような方法で教師チームプレイヤの行動選択確率$\pi^*(a|s)$の値を設定した：

① パス相手がa^*と異なる場合は 0.001。
② パス相手がa^*と同じだが、パスする場所が一定以上離れている場合は 0.2。
③ 上記以外の行動であれば正解行動a^*とみなし、①②の選択確率の合計値を 1 から減じた値とする。

もし、①～③の終了後、正解行動が行動候補$A(s)$にない場合は、①②の選択確率の合計を 1 から減じた値を$\pi^*(a^*|s)$とし、正解行動a^*を$A(s)$加えた。

5.　実験結果と考察

実験では、取得した 176490 個の学習データを、ランダムな 10 個のミニバッチに分割した。そして、順番にミニバッチごとの$\Delta\omega$を計算して平均をとり、重みを更新した。全てのミニバッチで重みを更新する作業を1エポックと定義し、50エポックの学習を行った。このときの損失関数であるカルバック・ライブラー情報量δ_{KLD}とエポック数の関係を図 1 に示す。

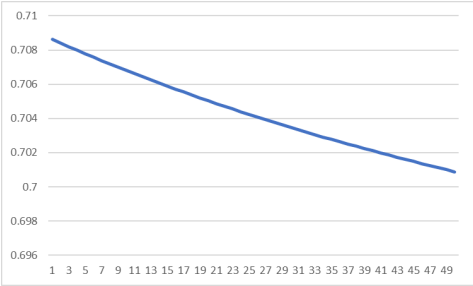

図 1　学習時の損失関数δ_{KLD}とエポック数の関係

学習前/後の評価関数をチェーンアクションに組み込んだチームと agent2d との試合を 100 試合行った結果が表 1 である。

表 1 対 agent2d の対戦結果

	勝率	勝ち	負け	分け	得点	失点
学習前	40%	35	52	13	1.70	2.23
学習後	50%	38	37	25	2.08	2.03

表 1 から、学習を行ったチームの方が学習前のチームより勝率、平均得点が上昇した。ただし、損失関数δ_{KLD}の減少が十分とは言えないため(約 1%)、学習エポックが足りておらず、さらなる学習が必要だと考えられる。

6.　今後の課題

本研究では、強豪チームの試合ログから評価関数の模倣学習を行った結果に、さらに教師あり学習を行う事で、改良を試みた。その結果、評価関数を組み込んだチームの勝率は、学習後の方が高くなったが、さらなる学習が必要であると考えられる。

現在は、学習を継続して行っており、学習が進んだところで、再度チームに組み込んで対局実験を行いたいと考えている。

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] 岡崎弘侑, 五十嵐治一, “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”, 情報処理学会第 83 回全国大会講演論文集, pp333-334, 2021

[3] 山岸拓海, 五十嵐治一, 山岸準, 入倉雅春, “サッカーエージェントの攻撃時における評価関数: 方策勾配を用いた教師あり学習”, 第 34 回ファジィシステムシンポジウム講演論文集, pp682-687, 2018

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.