ISS-P-259

連続空間における信頼度予測を用いた強化学習

蔡　詩祐†　　平原　誠†

† 法政大学大学院　理工学研究科

1. はじめに

強化学習を扱う上での課題の一つに，価値関数に基づく行動選択と，価値関数の推定値を改良する探索的行動とのバランスが挙げられる[1].

学習の進度が進むにつれて探索確率は下げていくのが理想であるが，問題に応じて学習進度の推移が異なるため，探索確率の適切な変更スケジュールを事前に設定するのは困難である. この問題に対して，”信頼度”という指標に基づいて探索確率を制御する提案がされている [2]. 本研究ではこの信頼度による探索確率の制御を連続空間における時系列課題に適用する.

2. 先行研究

2.1 離散空間上での信頼度による探索確率の制御[2]

離散空間におけるQ学習の一般式は

$$Q_{t+1}(s_t,a_t)=Q_t(s_t,a_t)+\alpha\delta_t \quad (1)$$

$$\delta_t=r_t+\gamma\max_{a'}\{Q_t(s_{t+1},a')\}-Q_t(s_t,a) \quad (2)$$

で表される. 式(2)はTD誤差δ_tと呼ばれる.

状態sに対する信頼度の逆数$R^2(s)$は

$$R^2_{t+1}(s_t)=R^2_t(s_t)+\alpha_R(\delta_t^2+\gamma_R R^2_t(s_{t+1})-R^2_t(s_t)) \quad (3)$$

で更新される. ここでα_R，γ_Rは0以上1未満の係数である. R^2値は，二乗TD誤差δ_t^2が大きい場合に増加し，小さい場合は減少することになる.

このR^2を用いて，状態sで行動aを選択する確率$P(s,a)$を

$$P(s,a)=\frac{\exp\left(\eta\frac{Q(s,a)}{R(s)}\right)}{\sum_{a'}\exp\left(\eta\frac{Q(s,a)}{R(s)}\right)} \quad (4)$$

により定める. softmax 法の温度パラメータ部分をRとすることで，学習の進度に応じて$P(s,a)$のランダム性を調整することを表している.

2.2 連続空間における時系列課題での強化学習[3]

連続空間における時系列課題に対して，強化学習を可能にするため，Elman 型リカレントネット(図 1)を用いてQ学習の価値関数近似を行う手法が提案されている[3]. 行動の数だけ出力素子を設け，各素子は入力(状態s_t)に対して行動価値関数Q^Nを出力する. 時刻tで選択した行動をa_tと表すと，その行動価値関数$Q^N_{a_t}(s_t)$に対する教師信号$T_{a_t,t}$は

$$T_{a_t,t}=r_{t+1}+\gamma\max_a Q^N_a(s_{t+1}) \quad (5)$$

で与えられる. ここでγは0以上1未満の係数，rは報酬である. この教師信号を用いて BPTT 法により学習を行う.

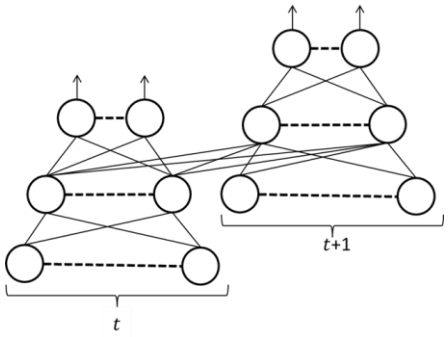

図 1 Elman 型リカレントネット

3. 提案手法

信頼度による探索確率の制御を連続空間上で実現するため，R^2値を，単出力の Elman 型リカレントネットで近似する. 出力をR^2値とし，R^2値に対する教師信号T^Rを

$$T^R_{s_t}=\left(\delta_t^N\right)^2+\gamma_R R^2_t(s_{t+1}) \quad (6)$$

$$\delta_t^N=T_{a_t,t}-Q^N_{a_t}(s_t) \quad (7)$$

により定める. $\left(\delta_t^N\right)^2$は，行動価値関数$Q^N$とその教師信号$T$の二乗誤差である. $\left(\delta_t^N\right)^2$が大きいと信頼度のネットワークの出力である$R^2$値は増加，小さい場合は減少する方向へ学習することになる. Rを式(4)の温度パラメータとして用いることで，連続空間上で信頼度による探索確率の制御が可能であると考えられる.

参考文献

[1] R.S. Sutton, and A.G. Barto, Reinforcement Learning: An Introduction, The MIT Press, 1998.

[2] 阪口豊,高野光雄，”内部モデルの信頼度に基づく強化学習のアルゴリズム,” 日本神経回路学会第 11 回全国大会予稿集, 2001.

[3] Kenta Goto, and Katunari Shibata, ”Eemergence of Prediction by Reinforcement Learning Using a Recurrent Neural Network,” Journal of Robotics, Volume 2010.