RF-005

確率環境下での優先度スイープ

Prioritized Sweeping in Stochastic Environments

鶴岡　久†　　田村剛士†　　山口明宏†

Tsuruoka Hisashi, Tamura Takeshi, Yamaguchi Akihiro

Model based learning can reevaluate the utility of every state, according to a measure of urgency. Prioritized sweeping is a typical algorithm for efficient state updating. In a stochastic environment, a probability distribution can be used to represent the uncertainty of the Q-value caused by probabilistic state transitions or probabilistic rewards. If the expected value of a reward is assumed to be normally distributed, the distribution of the value function at the initial learning stage is approximated by a t-distribution because of equivalence to random sampling from a normal distribution. Confidence intervals calculated from this distribution for each state-action pair represent insufficiently explored states.

In this paper, the product of the confidence interval and the Bellman error is used to provide a measure for prioritizing, which takes account of the level of confidence and also yields a measure of urgency. The performance of this approach in the trap domain is examined and compared with that of the ordinary sweeping method. Experimental results indicate that the proposed approach results in a more effective exploration of the state than does the use of conventional sweeping methods.

Keywords: Reinforcement Learning, Prioritized Sweeping, Confidence Interval, t-Distribution

Ⅰ. はじめに

強化学習には TD 学習のように実際の行動を通して経験から価値を推定し、方策を改善するモデルなし学習と、エージェントが観測結果から環境のダイナミックスを構築して状態価値を推定し、方策改善を行なうモデルベース学習がある。

未知の環境を不確定な行動をしながら学習するロボット等には確率環境下で学習効率の良い強化学習アルゴリズムが求められる。優先度スイープはモデルベース学習の代表的手法であるものの、確率環境下では確率論的モデルに適応させる必要がある。

Ⅱ. 背景

TD 学習は環境が作り出す実際の経験からバックアップ操作により価値の推定を行なう。これに対しモデル学習では、モデルが生成したシミュレーション上の経験も活用してプランニング手法により、学習を効率的に進めることが可能である。プランニングでは実際の経験から環境のモデルを構築し、価値関数と方策を改善する。代表的な方法として Dyna-Q があるが、状態数が多い大規模問題に対してはバックアップを緊急度の優先順位、通常は実際の経験から得られた価値関数とモデルで得られる価値関数の差（Bellman Error）の大きいものを優先する、優先度スイープが有利とされている[1]。

状態遷移や報酬獲得が確率分布で与えられる確率環境では、価値関数も学習過程では不確実性を有し、Bellman Error そのものが不確実性をもつため、決定論的環境と同じ意味で優先度を考えることができない。状態遷移や報酬が確率分布で表される確率環境下での価値関数の分布は R.Dearden 等によって考察され[2]、行動選択手法に応用されているが、優先度スイープへの直接的な応用はない。

Ⅲ. 提案手法

4 組(S,A,PT,PR)（ S:状態,A:行動,PT:状態遷移確率（状態 s で行動 a を起こしたときの状態 s' に到達する確率）,PR:報酬獲得確率（状態 s で行動 a を起こしたときの報酬 r を受け取る確率））で表現されるマルコフ決定過程（MDP）を考える。最適価値関数 V^*、最適行動価値関数 Q^*、は下記の Bellman 方程式に従う。

$$V^*(s) = \max_{a \in A} Q^*(s,a)$$

$$Q^*(s,a) = \sum r \Pr(r|s,a) + \gamma \sum_{PT(s \to s')} V^*(s')$$

状態sで行動 a を起こしたとき、状態遷移先s'がランダムに揺らいだり、獲得報酬がランダムなノイズを受ける確率環境下では価値関数の値やQ値が不確定な値になる。その分布関数は、学習過程において複雑であるが、訪問回数の増加にともない、価値関数の分布は正規分布に近づくことが示されている[3]。従って学習初期の訪問回数が少ない場合は、実現する価値関数の値は正規分布に従う母集団からのランダム抽出と等価と考えられるので、価値関数の分布は訪問回数を自由度とするt分布で近似できる。

M.Wiering [4],A.L.Strehl [5]等は状態遷移確率分布の信頼区間の上限を利用し、行動選択の探査効率の向上を図っている。しかし優先度スイープの優先度に直接関係するのは状態価値であるから、確率環境下での優先度を次のように考える。それまで訪問した価値関数の分布における信頼区間は、未知環境に対する新しい知見が得られる長期的尺度を表すと考えられる。また状態価値の信頼区間のみを優先度に反映させたのでは即応的な学習改善が図れないと考えられる。

そこでモデル価値と実測価値の誤差を表す従来の Bellman Error は報酬改善が期待される短期的尺度と考え、双方を考慮した優先度を採用する。本論文では両者の積に応じた優先度に従ってスイープする。Fig. 1 に提案手法の処理手順を示す。s1～sn は状態 s に対する現時点から過去に遡る n 個のサンプリング点を示し、これからt分布近似による信頼区間が計算されている。

Ⅳ. 実験結果

Fig.2に提案手法を評価するためのフィールドを示す。エ

† 福岡工業大学 Fukuoka Institute of Technology

```
Initialize V(s),Q(s,a),Model(s,a)
  Initialize s
  Repeat(for each episode)
        Repeat for each step of episode
       Q(s,a) ← Σ P(R+γ V(s' )),
          V ← max Q
       Choose a from s using policy from Q
          Next state s' ,
                Model(s,a)←s' , r
          Update V(s1) … V(Sn)
       Planning
          Calculation of Confidence Interval
          Decision of Prioritized sequence
          Repeat Ns times
              Q,V update
  until s is terminal
```

Fig. 1 Proposed Algorithm

ージェントはSから出発し、フラグを集めてゴールに入るとプラス1の報酬を得て1エピソードが終了するが、途中トラップに入るとマイナス10の報酬が加算される。

S	F	
		T
		G

Fig.2 Domain used in the experiment
S predicates the start space, G is the goal state, T is the trap state and F is the flag.

行動(up,down,left,right)は進行方向が空いていれば、確率0.9で移動が成功するが、確率0.1で希望する方向と直角方向へ動く。学習収束判定は続けて2回最短ステップでゴールしたこととする。Table 1に学習収束するまでのステップ数の50回平均値を示す。1エピソード毎に行なう優先度スイープはどの方法でも状態数に等しく9回とした。t 分布の信頼区間は両側 95%を採用した。

Table 1 Comparison of sweeping methods

Sweeping	Steps to Convergence
DYNA	14.66
Prioritized	10.58
Confidence Interval	10.76
Proposed	10.12

Confidence Interval は信頼区間の大きさのみを優先度の判断としたことを示す。Table 1から提案方式がBellman Errorのみを指標とした従来の DYNA スイープ、優先度スイープより学習効率が良いことが分かる。

次に 20 回学習を繰り返したときの、平均獲得報酬の変化をFig.3に示す。優先度スイープは学習初期でマイナス報酬が大きく、トラップを回避できていないことを示す。4 ステップ以降を拡大したのが Fig.4である。DYNA や優先度スイープは獲得報酬が容易に安定して正の値に向かわないことが分かる。

V 結言

学習変化していくモデルを活用して学習効率を向上させる方向には行動選択とプランニングがある。本論文は後者に注目し、確率環境下での優先度スイープに従来の1ポイントの時間に着目するBellman Errorによる優先度決定より、時間スケール

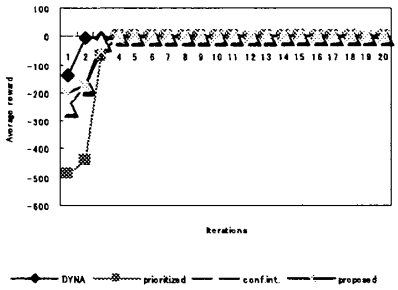

Fig.3 Plots of average reward as a function of number of iterations

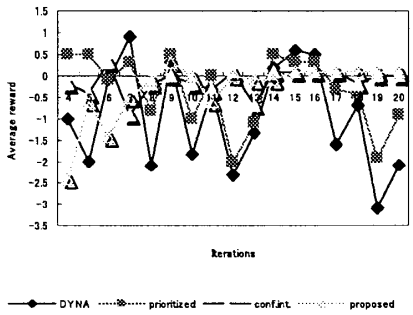

Fig.4 Enlarged plots from 4 to 20 iterations

の長い信頼区間も考慮した効率の良い優先度を提案し、実験的にこれを確認した。

行動から得られる経験から状態遷移分布や報酬分布を学習していく手段にはベイジアンモデルや母数モデルがあり、正確には信頼区間はこのような個々の経験によって個別に求められた分布から決めるべきと考えられるが、今後の課題である。

VI 参考文献

1)A.Moor and C,G,Atkeson: Prioritized sweeping;Reinforcement learning with less data and less time,Machine Learning,13,pp.103-130(1993)

2)R.Dearden, N.Friedman, S.Russsell : Bayesian Q-learning, Fifteenth National conf. On Artificial Intelligence (AAAI),pp.761-768(1998)

3) R.Dearden, N.Friedman, D.Andre : Model based Bayesian Exploration, Proc. Fifteenth Conf. On Uncertainty in Artificial Intelligence, (1999)

4)M.Wiering, J.Schmidhuber: Model-Based Exploration, From Animals to Animats 5,pp.223-228(1998)

5)A.L.Strehl, M.L.Littman : An Empirical Evaluation of Interval Estimation for Markov Decision Process, The 16th IEEE Int. Conf. on Tools with A.I.(ICTAI-2004),pp.128-135(2004)