高次元連続状態空間での強化学習における クリティカル状態を利用した適応的関数近似

二本 真 (510089)

北陸先端科学技術大学院大学 情報科学研究科 情報処理学専攻

2007 年 2 月 8 日

キーワード: 強化学習, NGnet, 二足歩行ロボット, 関数近似, クリティカル状態.

強化学習は機械学習のひとつで，学習主体 (エージェント:agent) が環境との相互作用によって最適な行動系列を習得するための手法であり，幅広い分野で応用されている．強化学習では行動の良し悪しを判断する教師は必要とせず，代わりに，設計者が予め目標の状態に設定する報酬を便りに学習する．これにより，設計者にとって未知な環境でも報酬を設定 (報酬関数の定義) するだけで，エージェントが最適な動作を試行錯誤により自動で獲得できるという利点を持つ．しかし，探索空間が巨大になると，計算機メモリを大量に消費するだけでなく，適切な行動系列を獲得するまでに非常に時間がかかる．また，強化学習は現状態から今後得られるであろう報酬の期待値を状態の評価値とし，評価値に基づいて行動選択を行っているため，報酬までに長い行動系列が必要となる場合，適正な評価値を得るまでに多くの試行錯誤を要せねばならない (報酬の遅れ)．遅れが小さい報酬の設定が理想であるが，対象のタスクについての知識が必要となり，一般に難しい．こうしたなかで，より効率の良い学習アルゴリズムが求められている．従来の強化学習では，どのように状態空間を制限し，学習効率を向上させるかが大きな位置を占めていた．例えば，石井らは 2 足歩行をリズム運動であると仮定し，両手，両足，腰などの運動が同期するとして 状態探索空間を限定し，2 足歩行の強化学習を行った．しかし，この状態空間制 限は 2 足歩行にのみに有効であり，他の問題に応用できるものではない．このように，従来の強化学習では，問題に特化した方法を用いて探索空間を制限しているが，自律的な学習という強化学習本来の利点が損なわれるため，問題の性質に依存した手法は好ましくない．本研究では，問題の性質に依存せずに，より高次元な状態空間を持つ問題に対して適用できる，ゼロベースでの強化学習システムの構築を目指す．

強化学習では，方策関数や状態価値関数を近似し，それらを徐々に変更していくことで学習を行う．方策関数とはエージェントの行動出力を決定する関数で，状態価値関数とは，ある状態から未来においてどれだけの報酬が期待できるかを示す関数である．強化学習でよく用いられる近似手法のひとつに，正規化ガウス関数ネットワーク (normalized

Gaussian network:NGnet) があげられる．NGnet は，ユニットと呼ばれる基底関数を状態空間に配置して，その発火量を用いて関数の近似を行う手法で，状態空間に配置されたユニットは，エージェントがユニットの中心に近い場所にあれば強く発火し，中心から遠い場所にあれば弱く発火するので，滑らかな曲線での近似が可能である．しかし，NGnet で状態空間が非常に大きなものの近似を行おうとすると，その状態空間を埋め尽くす数のユニットが必要になるので，大量の計算機メモリが必要になるという問題がある．もし，状態空間の大きさに見合った数のユニットを配置しないで近似を行う場合は，精度の悪い結果しか得られない．これにより，強化学習のにおいて，方策関数，状態価値関数の近似精度は，学習の精度自体に大きく関わる重要な要素なので，学習する問題の状態空間が大きいと，必然的に学習自体の精度も悪くなってしまうのである．そこで，本研究では，ユニットをより効率的に用いることで，状態空間の大きさに対して少ないユニット数でも，関数近似の精度を悪化させず，安定した学習結果が得られる手法を提案する．本研究では，ユニットを効率的に用いる手段として，ユニットが配置されている状態空間上の位置に着目した．NGnet を用いた関数近似において，近似誤差は状態空間上の全ての場所で一定ではなく，近似誤差が大きい場所や小さい場所があると考えられる．また，状態空間が大きな問題では，エージェントがあまり通ることのない状態が多く存在しており，そのような場所に配置されたユニットは，学習中にほとんど発火することがないと考えられる．それらのことから，状態空間には，ユニットを配置することで学習のために有効である場所とそうでない場所があると考えられ，学習に有効な場所だけにユニットを集中させることができれば，無駄に使われるユニットがなくなり，少ないユニット数でも安定した学習結果が期待できる．

従来の手法を用いてポールバランシング問題を対象に予備実験を行った結果, エージェントの挙動が安定するか不安定になるかのターニングポイントであるクリティカル状態が確認された. 本研究ではこのような場所を, 学習中のサンプルをもとに自動的に検出し, ユニットを配置することで，学習の安定化を図った．また, 実際に強化学習システムを実装し，予備実験より広大な状態空間を持つ二足歩行問題を対象として実験を行った．その結果, 従来手法と比較して, たかだか 44.06 %のユニットで同等以上の学習精度を実現し, 提案手法の有効性を確認した.