様式 C-19

科学研究費補助金研究成果報告書

平成２１年６月１９日現在

研究種目：基盤研究（C）
研究期間：2007～2008
課題番号：１９５００１７２
研究課題名（和文）　次元圧縮のできる多変量解析手法を用いた
強化学習エージェントの性能解析
研究課題名（英文）　Performance Analysis of a Reinforcement Learning Agent Using Multivariate Analysis Method Based on Dimension Reduction

研究代表者
釜谷　博行（KAMAYA HIROYUKI）
八戸工業高等専門学校・電気情報工学科・教授
研究者番号：70224657

研究成果の概要：高次元の連続状態空間を直接扱うことのできるモデル追加型の強化学習アルゴリズムを開発した。このアルゴリズムの最大の特徴は、関数近似器のパラメータをうまく設定することで､モデル数を小さく抑えつつも良好な学習性能を実現できる点にある｡このため、まず、パラメータの挙動解析を行い、最良のパラメータを見出した。つぎに、10 次元の連続状態空間をもつ移動ロボットの移動障害物回避問題に適用し、有効性を確認した。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	1,900,000	570,000	2,470,000
２００８年度	1,500,000	450,000	1,950,000
年度			
年度			
年度			
総　計	3,400,000	1,020,000	4,420,000

研究分野：総合領域
科研費の分科・細目：情報学・知覚情報処理・知能ロボティクス
キーワード：強化学習、関数近似、自律移動ロボット

１．研究開始当初の背景

将来、人間と共存し、人間の代わりとなって働くようなロボットを実現するためには、固定された制御ルールを用いるだけでなく、動的に変化する環境の中で、ロボット自身が学習によって制御ルールを獲得することが要求される。そのような要求に応えるため、未知環境においてロボットに行動を獲得させる手法として注目を集めているのが強化学習である。

強化学習を実問題へ適用するには、連続な状態空間を取り扱わなければならない。連続状態空間に対するアプローチとして、小脳の計算モデルである CMAC（タイルコーディング）、階層型ニューラルネットワーク、動的基底関数（RBF）などの関数近似を用いる方法が提案されている。しかし、これらの関数近似法は、ノイズへのオーバーフィッティングなどによって近似がうまくいかないなどの失敗例が報告されている。また、比較的低次元の問題への適用例しかなく、高次元問題への拡張が難しいと考えられる。

２．研究の目的

本研究の目的は、実用化を図る上で重要となる高次元連続状態空間の問題に対して、次元圧縮可能な多変量回帰分析を用いることで、膨大な情報の中から報酬に結び付く重要

な情報のみを自律的に抽出できる汎用性の高い強化学習システムを実現することにある。本研究では、多変量解析手法の一種である局所重み付き部分最小二乗法(LWPLS: Locally Weighted Partial Least Squares)を関数近似手法とする新たな強化学習アルゴリズムを提案する。学習精度を確保しつつ、リアルタイム性能を向上させ、最終的には高次元の連続状態空間を扱う問題に適用し、提案手法の評価を行う。

３．研究の方法

(1) 多変量解析のためのサンプルデータの採取方法と Q 値の更新則の検討

多変量解析を行う上で必要なサンプルデータの効率的な取得方法について検討する。従来のアルゴリズムでは、状態空間のサンプルデータは学習前に格子状に予め用意しておき、学習時には Q 値のみを更新していた。この方法では、高次元空間を扱う場合にメモリ不足によりプログラムを実行できないなどの問題が発生する。

ここでは、学習時にある条件を満たした場合にオンラインでサンプルデータを適宜追加していく方法について考案する。また、学習性能の向上を目指した Q 値の更新則についても検討する。

(2) パラメータの挙動解析

センサ情報が得られてから行動決定までのレスポンスを向上させるため、学習性能を低下させることなく記憶するモデル数をできるだけ少なくするようなパラメータを見出す必要がある。これらの観点からパラメータについて再検討する。

学習を成功する上で重要なパラメータはどれか、また、パラメータの値はどのように決めるべきかなどの指標を獲得することを目的として、提案システムにおいて各種パラメータを変更した場合の学習性能について詳細に調べる。

(3) 時系列情報を利用した場合の性能解析

部分観測問題に対しては、現在のセンサ情報に加えて、過去のセンサ情報の履歴を用いる方法に着目する。時系列情報を用いると高次の連続状態空間となるが、提案システムはある種の部分観測問題に対してうまく学習できるかについて検証する。

(4) OpenMPによる並列計算機環境の構築と提案システムの並列化

多変量解析における計算量を実用レベルまで下げ、実システム制御時のリアルタイム性を確保するため、Linux ベースの並列計算機を用いて実行環境を構築する。つぎに、並列計算を行うため OpenMP に準拠したコンパイラを新たに導入するとともに、提案システムを並列計算機上に実装し、その実時間性能について評価する。

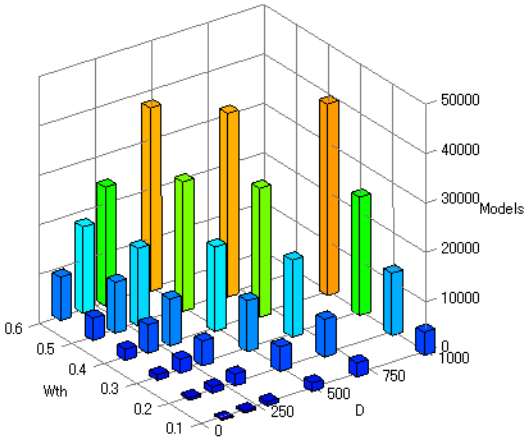

図１　最終試行までに追加されたモデル数

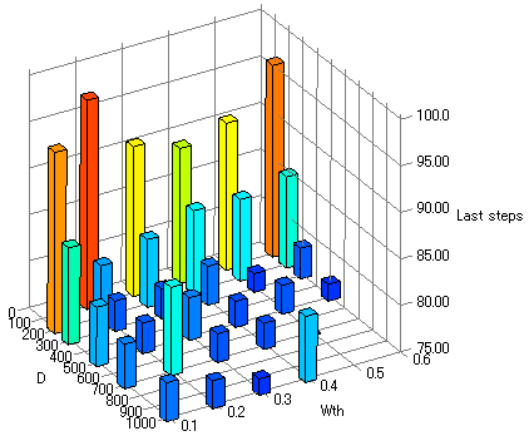

図２　最終試行におけるゴールまでのステップ

(5) 高次元連続状態空間問題への適用

これまでの研究では、2 次元あるいは 4 次元の比較的低次元の連続状態空間をもつ問題において提案手法の有効性を確認してきた。ここでは、10 次元という高次元の連続状態空間をもつ移動ロボットの移動障害物回避問題に提案手法を適用し、シミュレーション実験により、その有効性を評価する。

４．研究成果

(1) オンラインでのサンプルデータの追加方法をつぎのようにした。まず、状態空間において、推定点からのユークリッド距離に基づき各サンプル点での重み w_i を計算し、それらの中の最大値 w_{max} を求める。つぎに、w_{max} が閾値 w_{th} 未満の場合に、推定点をサンプルデータとして追加する。

また、学習性能の向上を目指した Q 値の更新法には適格度トレースを利用した。具体的には、適格度トレースパラメータの更新に上

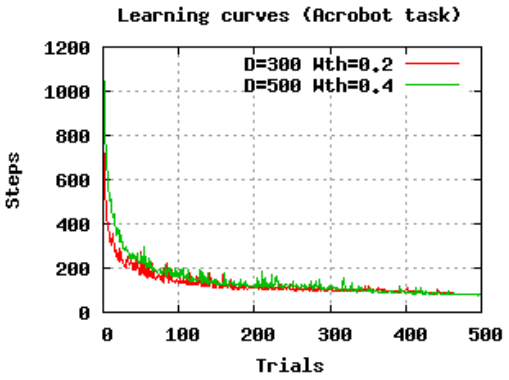

図３　学習曲線（ゴールまでのステップ数）

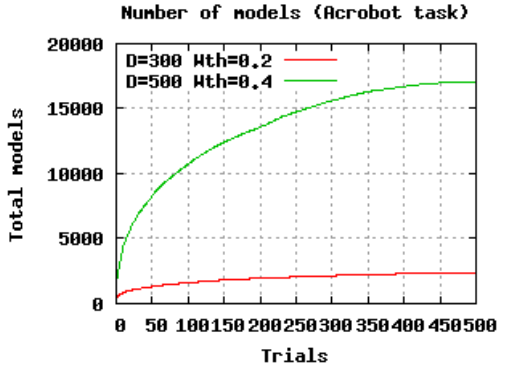

図４　試行回数に対するモデル数の変化

述の各サンプル点での重みw_iを用いた。各サンプル点におけるQ値の更新にはSarsa(λ)学習を適用し、極めてシンプルな形で学習アルゴリズムを実装した。

(2) 提案手法において重要なパラメータは、重みw_i計算時に用いられる係数Dとサンプルデータ追加時の閾値w_{th}である。これまでの予備的な実験では、D=500, w_{th}=0.4 が最良な値として認識されていた。

ここでは、アクロボット問題を用いて、Dとw_{th}を変化させたときの挙動解析を行った。アクロボット問題とは、2 リンクのアームをある高さまで振り上げるという強化学習において代表的な問題の一つである。評価は、最終試行までに追加されたサンプルデータのモデル数（図１）と最終試行におけるゴールまでのステップ数（図２）の２つの指標で行った。

図１の結果をみると、Dまたはw_{th}が増えるにしたがい、モデル数が増加することがわかった。これに伴い、計算時間の増大が確認された。一方、図２の結果をみると、w_{th}にはあまり依存しないが、Dが小さくなるにしたがい、ゴールまでのステップ数が増加し、学習性能が低下するということが確認できた。以

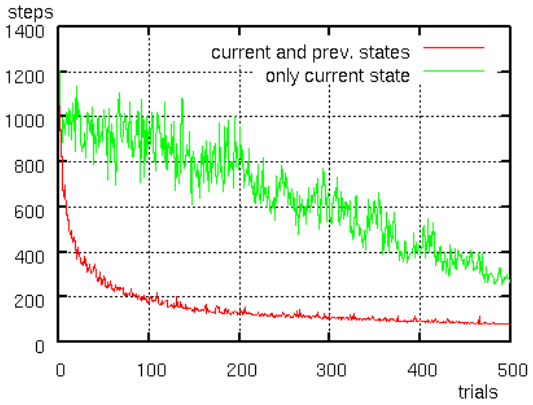

図５　学習曲線（ゴールまでのステップ数）

上のことから、モデル数を低く抑えるには、w_{th}をあまり大きくできない。また、良好な学習性能を維持するには、Dを小さくし過ぎないようにに適切な値に設定しなければならないということがわかった。結果として、Dおよびw_{th}として、それぞれ 300, 0.2 付近の値が最良であることが確認できた。

このことを検証するために、下記に示す 2 種類のパラメータ設定について評価した。

設定 1：D=500, w_{th}=0.4

設定 2：D=300, w_{th}=0.2

図３は、試行回数に対するゴールまでのステップ数の変化（50 回のシミュレーションの平均値）を示す。両設定ともに、全体的にほぼ同じ特性を示した。学習初期を比べると、設定 2 の方が、設定 1 に比べて若干速く収束していることがわかる。最終試行でのステップ数は、設定 1 では 77.9、設定 2 では 78.3 とほぼ同等となった。

図４は、試行回数に対するモデル数の変化を示す。学習初期はモデル数が急激に増えるが、後半は傾きが緩やかになり、最終的には収束傾向にあることがわかる。最終試行でのモデル数は、設定 1 は 17,066 であったのに対して、設定 2 では 2,349 となり、大幅に軽減できた。モデル数が少ないと計算時間が短くなり、学習器のリアルタイム性能を向上できる。このことは、より高次元の問題に提案手法を適用する上で重要となる。

これ以降の実験結果は、すべて設定 2 のパラメータを用いて行ったものである。

(3) アクロボット問題において、2 つのリンクの現在の角速度情報が得られないものとして、部分観測問題を考える。この問題に対応するため、リンク角度の時系列情報を利用する。実験では、現在の角度情報のみを用いた場合（2 次元入力）と、現在の角度情報と 1 ステップ前の角度情報を用いた場合（4 次元入力）について比較・検討した。

図５に学習曲線（50 回のシミュレーションの平均値）を示す。現在の状態のみでは、学

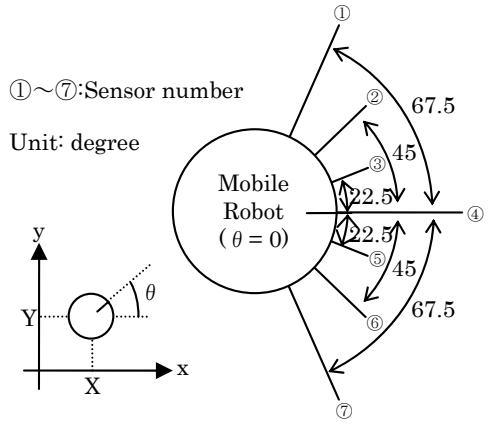

図６　距離センサの計測方向

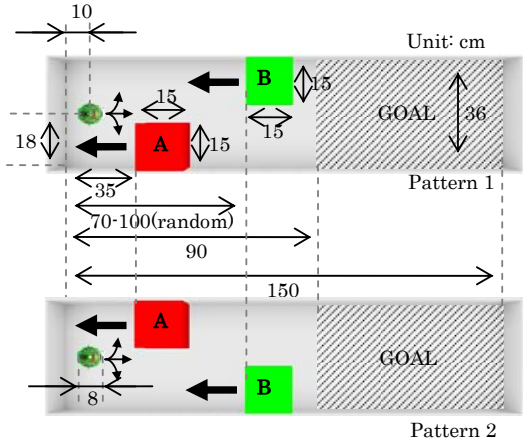

図７　実験環境

習曲線が振動的となり、うまく学習できなかった。これに対して、1 ステップ前の情報を用いると良好な収束特性を示し、最終試行でのステップ数は 79.2 であった。時系列情報を利用することで、提案手法は部分観測問題へも適用可能であることが確認できた。

(4) クワッドコア CPU を 4 個搭載した計算用サーバー(コア数は 16)を構築し、OpenMP 対応のコンパイラである Intel C++を導入した。

アクロボット問題において、関数近似手法により Q 値を推定する部分を行動毎に並列化した。行動数が 3 および 16 について、並列化しない場合と並列化した場合の処理時間を比較したものを表 1 に示す。結果として、並列化により実時間性能が向上することを確認できた。特に、行動数の大きい方が並列化の効果が顕著（速度比で 7.44 倍）に表れていることがわかった。

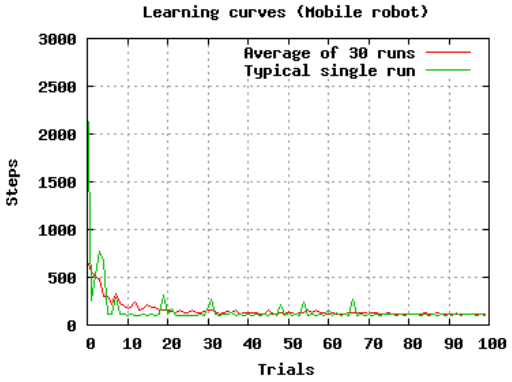

図８　学習曲線（ゴールまでのステップ数）

表１　処理時間の比較

行動数	処理時間[sec]		速度比
	非並列	並列化	
3	43.0	22.2	1.94
16	335.0	45.2	7.44

(5) 移動ロボットがスタート位置から移動障害物を避けてゴールを目指すという問題において、提案手法を評価する。静止障害物の回避と異なり、移動ロボットの位置情報に加えて、障害物までの距離情報を取得する必要がある。ここでは、測域センサを搭載した移動ロボットを想定してシミュレーション実験を行う。

移動ロボットは自己位置と進行方向の推定が可能であると仮定し、位置 x[cm]∈[0, 150]、 y[cm]∈[0, 36]、x 軸と進行方向とのなす角度 θ [rad]を取得できるものとする。また、測域センサを仮定した距離センサから、進行方向とそれを中心に 22.5[deg]毎に左右それぞれに 3 方向、計 7 方向の距離を最大 50[cm]まで取得できるものとする(図６)。これらの入力により 10 次元の連続状態空間を構成する。ロボットの初期位置(x, y)= (10, 18)で、θ =0 とする。ゴール条件は x > 90 である。行動は前進、右折、左折の 3 つとし、車輪の速度で定める。今回用いた設定は、前進が(10, 10)、右折が(10, 2)、左折が(2, 10)である。カッコの左側の値は左車輪の速度、右側の値は右車輪の速度を表し、単位は[cm/s]である。報酬はゴール到達で+30、障害物との衝突で－30 を与え、その他は 0 とする。衝突時にはスタート位置から再スタートする。移動ロボットの直径は 8[cm]である。

ゴールに到着すると 1 回の試行が終了する。1 回の試行の最大ステップ数は 3,000 で、1 回のシミュレーションで 100 試行を行う。最大ステップ数に達するとゴール到達ができなくても次の試行を開始する。

移動障害物の初期配置は図 7 に示すよう

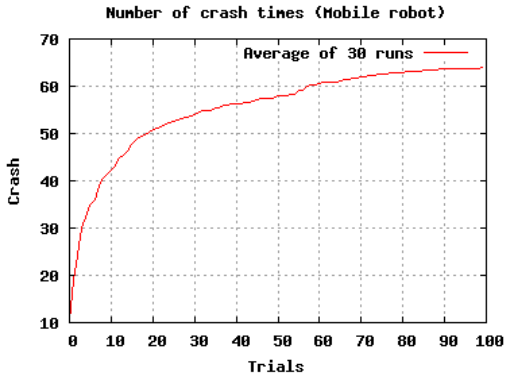

図９　学習曲線（衝突回数の累計）

に、２つのパターンとした。偶数番目の試行ではパターン１を、奇数番目の試行ではパターン２を用いた。移動ロボットに対して手前にある障害物 A は x=35 の位置に置かれ、奥の障害物 B は x=70~100 の範囲で、試行開始時にランダムに置かれる。これらの障害物は試行開始とともに-x 軸方向に約 2[cm/s]の速度で移動する。移動ロボットが再スタートする場合は、移動障害物はその試行時の初期位置に戻されるものとする。

ロボットシミュレータを用いて学習実験を行う。行動選択のサンプリング時間を 0.1[s]とし、行動はその間継続させる。シミュレーションの物理学計算は 0.02[s]毎に行い、床との摩擦やロボットの加速特性等を考慮した。

図８にゴールまでのステップ数の変化（30 シミュレーションの平均値と代表的な１シミュレーションの結果）を示す。また、図９に衝突回数の累計を示す。グラフを見ると、20 試行あたりから衝突回数が減少傾向となり、ゴールまでのステップ数も小さく安定し、良好に学習していることがわかった。最終試行は 116.8 ステップで、衝突回数の累計は 63.9 回となった。なお、このときのモデル数は 3,306 であった。１行動当たりの計算時間を算出すると約 0.04[s]となり、行動選択のサンプリング時間の 0.1[s]より小さくなった。このことから、提案手法は実機で動作させることが可能であるという知見が得られた。モデル数を小さく抑えることは、学習器の実時間性能を維持する上で重要であるといえる。今後は、実機上で動作検証を行う予定である。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計２件）

① 釜谷博行、藤村敦子、工藤憲昌、阿部健一：関数近似手法を用いた強化学習アルゴリズム、八戸工業高等専門学校紀要、43 号、pp. 23～27、2008、査読有
② 釜谷博行、 阿部健一：連続状態空間のための強化学習アルゴリズム、八戸工業高等専門学校紀要、42 号、pp. 65～68、2007、査読有

〔学会発表〕（計４件）

① 一井宏次、釜谷博行、工藤憲昌：高次元連続状態空間における強化学習－局所重み付き回帰手法を用いた価値関数近似－、計測自動制御学会東北支部第 250 回研究集会、2009. 6. 19、八戸工業高等専門学校
② 一井宏次、釜谷博行、阿部健一：局所重み付き回帰手法を用いた強化学習、平成 21 年電気学会全国大会、2009. 3. 19、北海道大学
③ 一井宏次、釜谷博行：強化学習のための局所重み付き回帰手法を用いた価値関数近似、平成 20 年度電気関係学会東北支部連合大会、2008. 8. 22、日本大学工学部
④ 工藤憲昌、田所嘉昭：適応周波数推定法の検討とその一応用、計測自動制御学会東北支部第 236 回研究集会、2007. 6. 15、八戸工業大学

６．研究組織

(1) 研究代表者

釜谷　博行（KAMAYA HIROYUKI）
八戸工業高等専門学校・電気情報工学科・教授
研究者番号：70224657

(2) 研究分担者
なし

(3) 連携研究者

工藤　憲昌（KUDOH NORIMASA）
八戸工業高等専門学校・電気情報工学科・教授
研究者番号：40270194