連続な感覚入力の強化学習

：多段の内部構造モデルによる接近

東京工業大学　総合理工　○佐藤誠　木村元　小林重信

Reinforcement Learning for Continuous Input using Multi Step Reinforcement Learning Model

○ Makoto SATO　Hajime KIMURA　Sigenobu KOBAYASHI
Department of Intelligence Science,Tokyo Institute of Technology

A bstract : To solve high dimentional continuous state space problems, we propose a new reinforcement learning method that has two step Reinforcement L earning layers. One layer is an action selector that stochastically selects an action, and the other is a state analizer that stochastically partitions the input space. E ach layer adapts itself according to the reinforcement signal. The proposed state analyzer can create new needed inner states. Computer simulations have been conducted to illustrate the performance and applicability of the proposed learning method.

1．″ ¶

強化学習は正解についての知識を必要とせず、環境中での試行錯誤で得られた報酬から環境へ適応する学習の枠組である。正しい出力値が必要ない点、行動系列を評価すれば良い点が強化学習の特徴であり，非線形性の強い制御問題、大規模なゲーム、複数のエージェントが 強調して問題を解決するマルチエージェント系の学習など、システムの正しい出力は設計者には分からないが、出力を評価することが可能な問題領域で有望と考えらる．大規模・複雑な実問題は環境の状態を識別するために多くの感覚入力情報を必要とするため，多次元・連続値の感覚入力を扱うことの可 能な強化学習手法が求められている。

ダイナミックプログラミングを基礎とする強化学習アルゴリズムであるTD 法 [Sutton 88] やQ-learning [Watkins et al 92]は lookup tableを用いる場合には最適な行動を 学習することが可能だが，実問題のような多次元・連続値問題に対してエージェントの内部状態空間表現にlookup tableを用いようとすると状態数の組み 合わせが爆発してしまう．そこで、ニューラルネットのような関数近似とValue Iterationを組み合わせた方法が用いられている[Lin 93], [Tan91], [Tesauro 92] , [Satinder 97], [Robert 96], [Mahadevan 97]しかし、関数近似を用いた場合必ず正しい値に収束するとは限らないことが報告されており[Boyan 95] , [Leemon 95] 問題によらず安定して学習可能な手法が求められる．

非Bellman型の強化学習手法である確率的傾斜法 [Williams 92], [Kimura 95]は、 政策の最適性は保証されないがlookup tableを前提としておらず、連続値入力問題に対して関数近似システムを組み合わせても比較的安定して局所的な合理性を満足する政策を学習可能である．そこで，連続値入力問題に対してロバストな学習能力が期待されるが，関数にどの程度の近似能力を与えるべきかの決定に設計者の試行錯誤が必要となるのが現状である．

実問題のように大規模・複雑な環境に対して適応可能な強化学習システムを構築することが本研究の目的である。そのような問題では関数近似システムを用いた汎化を行なう必要があり，その際重要となるのが、 関数近似システムにどの程度の近似能力を与えるべきかという問題である。 近似能力が不足するとエージェントは正しい行動出力を学習できなくなり、必要以上の近似能力を与えると、学習に膨大な 試行が必要となる。

そこで，学習中に近似能力を追加する事が可能な近似システムが有効と考えられる。本論文では、報酬から 適切な領域分割を学習し、さらに必要に応じて近似資源を投入していく関数近似手法を提案する。ここでは、連続値空間を適切に分割するために，センサ入力からエージェントの内部状態への関数を決定する状態認識器を学習の対象と考える． 状態認識器の出力を内部状態集合への確率ベクトルとし，確率的傾斜法の更新式を用いて状態認識器のパラメータを更新する． また，現在存在する内部状態のうちどれも選ばれない場合，適切な内部状態が存在しないと判断し新たな内部状態をつけ加える．低次元の連続値問題に提案手法を適用した計算機シミュレーションの結果から提案手法の振舞いおよび有効性を示す。

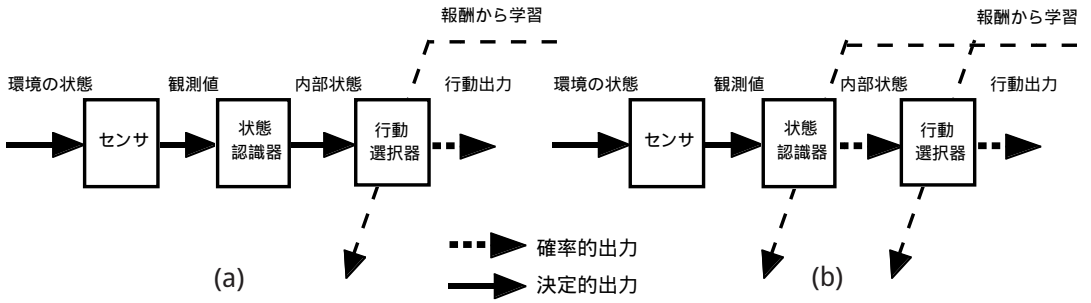


図1 ブロックダイアグラム(a:従来システム,b:提案システム)

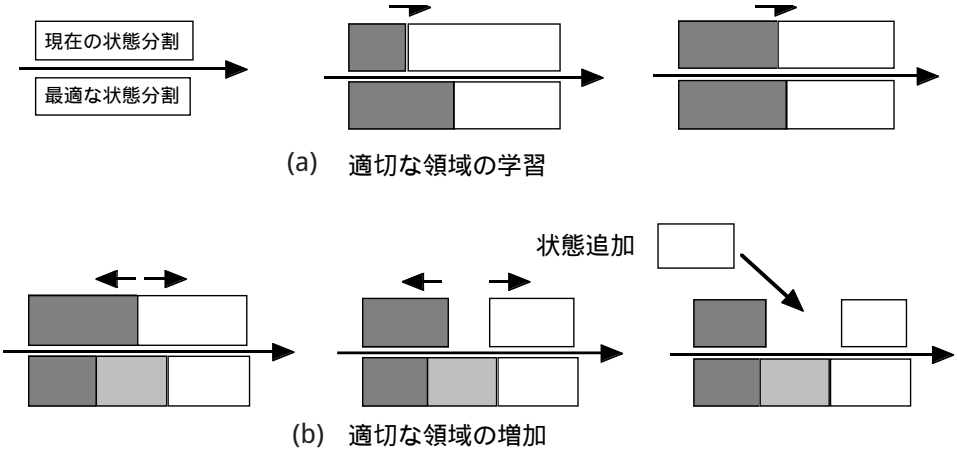


図2 状態認識器の学習

2．　Ł　　Ł ¤　　@

本研究では連続値入力・離散値出力の強化学習問題を対象にする．連続値入力・離散値出力問題では、観測値空間は適切な行動の異なる複数の領域に 分かれている．

強化学習エージェントの目的は獲得報酬の最大化である．適切な領域分割とその領域内での適切な政策の学習が行われた場合に、獲得報酬は最大となる．強化学習の設定では，学習中にある観測点での正しい行動が環境から直接与えられることは期待できない．また、異なる領域に移動した・しないという情報も期待できない． エージェントは様々な観測点で様々な行動を出力することを繰り返し、報酬をもとに環境が求める行動を学習することが求められる．さらに、相互作用を何度か繰り返した後に報酬が与えられた場合にも、適切な行動 系列を学習できなければならない．

特に，大規模な連続値問題では学習を行う場合、学習システムは以下のような条件を満たすことが求められる． 1,入力が連続値であるため、ある観測点での経験をその観測点に近い点での行動選択に反映させなければならない．エージェントが自ら観測空間を分割し、その領域内での経験を共有することにより領域内で求められる行動を学習するのに加え、自らの領域分割を実際の環境の適切な領域分割に近づける必要がある．2， 適切な分割領域数(近似資源)を決定しなければならない．近似資源が不足するとエージェントは適切な行動を学習不可能となる．資源が過剰になるにつれ学習に膨大な試行が必要となる．3,入力値が存在しない点のために近似資源を供給しない．

図1(a)は、一般的な強化学習システムのブロックダイアグラムである．第一のブロックはエージェントのセンサであり環境の状態から観測値への関数を与える． 第二のブロックは状態認識器であり観測値からエージェントの内部状態への関数を与える．第三のブロックは行動選択器であり内部状態から行動集合への関数を与える．行動選択器は学習の対象となり報酬から学習を行う．提案手法では状態認識器も学習の対象と考える(図1(b))．

連続値、高次元の観測入力から適切な行動を学習するためには状態認識器の学習が有効であると考えられる．状態認識器の学習では適切な状態空間の分割を行うことが必要である(図2(a))．また，新たな内部状態をつけ加えるための学習も必要である(図2(b))．本研究ではこのような学習を報酬を基に行う手法を提案する．

3.学習システムの提案

3.1.提案システムの概要

学習システムが環境から観測値xを観測すると，従来の手法では，何らかの決定的関数に従い状態認識器が全内部状態の中の 1 つを選択する．選択された内部状態の政策をもとにして，行動選択器が何らかの確率的関数に従い全行動の中の 1 つを選択し，環境に対し出力する．環境は行動を評価し，その評価値を報酬として学習システムに入力する．学習システムは報酬をもとに政策の学習を行う．

状態認識器は観測値空間の分割を行っている．提案手法では，分割を行う関数を確率的関数とし，調整の対象となるパラメータWを導入する．報酬をもとにして，パラメータWを確率的傾斜法の更新則に従い更新する．

観測入力x に対して，すべの内部状態について適合度M を計算し，M をもとに各内部状態i の選択確率を式(1) のように計算する．さらに，状態認識器が既存の状態を選択しない確率を式(2) に従い計算する．最終的に，システムの行動 aの選択確率は式(3)によって計算される．ここで，Tは行動選択器の学習対象となるパラメータである．

式(3) に確率的傾斜法のアルゴリズムを適用することにより，パラメータW ，および，T を更新する．さらに，必要に応じて新たな内部状態を追加する．

$$\Pr ob(i \mid x, W) = (1 - \prod_s (1 - M_s)) \frac{M_i}{\sum_s M_s} \quad (1)$$

$$\Pr ob(other \mid x, W) = \prod_s (1 - M_s) \quad (2)$$

$$\Pr ob(a \mid x, W, T) = \Pr ob(a \mid other) \Pr ob(other \mid x, W) + \sum_s \Pr ob(a \mid s, T) \Pr ob(s \mid x, W) \quad (3)$$

3.2.学習則

図3 に確率的傾斜法[Kimura 95]の学習アルゴリズムを示す。この学習法は政策の最適性は保証されないがlookup tableを前提としておらず、関数近似システムを組み合わせても報酬の改善を最大化する政策(パラメータ)を学習可能 である．図3式(4)の

1.環境の観測 X_tを受けとる
2.$\pi(a_t, W_t, X_t)$ の確率で行動a_t を実行する
3.環境から報酬 r_t を受けとる
4.内部変数W の全ての要素について$e_i(t)$ $\overline{D}_i(t)$ を求める ただしγ は割引率 $(0 \le \gamma < 1)$

$$e_i(t) = \frac{\partial}{\partial w_i} \ln\{\pi(a_t, W_t, X_t)\} \quad (4)$$
$$\overline{D}_i(t) = e_i(t) + \gamma \overline{D}_i(t-1)$$

5.以下の式を用いて $\Delta w_i(t)$を求める

$$\Delta w_i(t) = (r_t - b)\overline{D}_i(t)$$

ただしb は定数である

6.政策の改善：以下の式で更新

$$\Delta W(t) = (\Delta w_1(t), \Delta w_2(t), ..., \Delta w_i(t), ...) \quad (5)$$
$$W \leftarrow W + \alpha \Delta W(t) \quad (6)$$

ただしα は非負の学習定数である

7.時間ステップ t を$t+1$へ進めて1へ戻る

図3 確率的傾斜法アルゴリズム

左辺はタイムステップ　t におけるパラメータのeligibility [Singh 94]と呼ばれ、これは選択された行動による情報ゲインの ような量である。Eligibilityが計算可能ならば、図3の式(5),(6)に 従いパラメータの更新が可能である。そこで，提案手法では，式(7),(8)を図中の式(4)に置き換えることにより学習を行う．

$$e(t, W) = \frac{\partial}{\partial W} \ln(\Pr ob(a \mid x, W)) \quad (7)$$

$$e(t, T) = \frac{\partial}{\partial T} \ln(\Pr ob(a \mid x, T)) \quad (8)$$

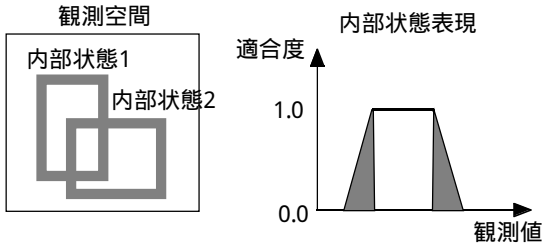


図4 内部状態表現

3.3.内部状態表現

各々の内部状態は図4のように状態空間の連続する局所的な領域をカバーするような適合度関数でなければならない。新たな内部状態を加えた影響をその新しい状態の周辺のみにとどめるためである。関数は図4 のように、適合度1 の領域、0 の領

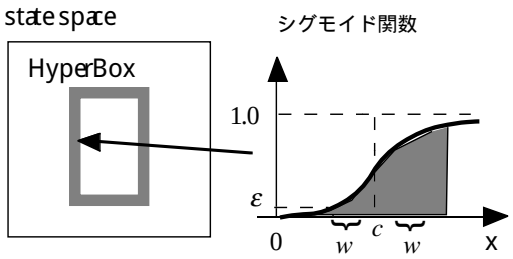


図5 HyperBox状態表現

域、0 ~1 の値を取る確率的な領域からなり、確率的な領域に観測値が入力した場合に学習が行われる。確率的な領域の適合度関数は内部状態のパラメータWで微分可能でなければならない。

そのような内部状態の一例として図5のような箱型の内部状態表現が考えられる．この表現では，箱の境界線が確率的な適合度関数になっている．この内部状態表現を用いた場合，次元数× 4 のパラメータが必要になる．

3.4.状態の追加

観測値が与えられたときに既存の内部状態を選択する確率が非常に低い場合、その観測値の周辺には内部状態が存在しないか、学習によってその領域をカバーしない 方が報酬に貢献できると判断したかのいずれかである。いずれにせよ、その観測点 周辺に新たな内部状態を追加する価値があると考えられる。そこで、式(9)の条件を 満たす観測点が存在する場合、新たな状態を追加することを決定する。 ただし、ε は十分小さな正の実数値でユーザが決定するパラメータ となる。例えば、図6のような内部状態の関数が存在し、観測点x1，x2，x3が入力された場合、観測点x1, x2に追加するよりも観測点x 3 への追加が優先される。各々の内部状態に関して観測値x2, x3での適合度は等しいが、 式(2)にはすべての状態の適合度が反映されるためである。

$$\Pr ob(other \mid x, W, T) > 1.0 - \varepsilon \qquad (9)$$

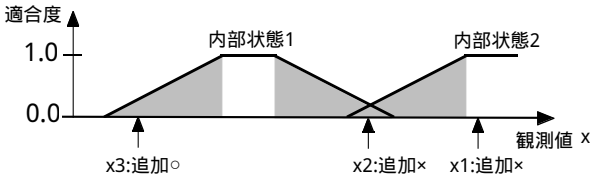
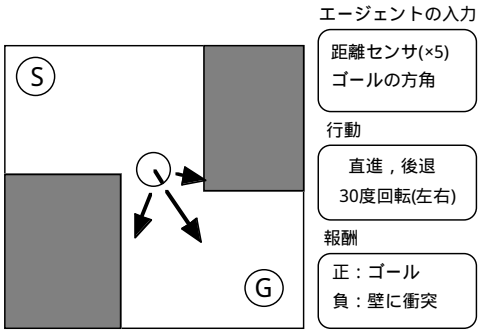


図6 状態追加



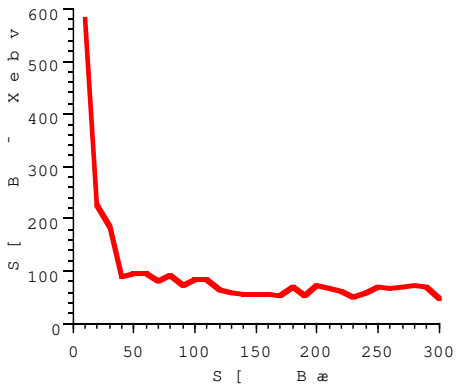


図7 ゴール到達問題，結果

4．　－

提案手法の有効性を示すため、 2 次元入力のPuddle World問題、および、6次元入力問題であるゴール到達問題に適用した。なお、内部状態表現はHyperBox表現を用い、新たに追加する状態の初期パラ メータは中心点として観測点を、確率的な適合度の領域の範囲はあらかじめ与えた固定の値とした。

4．1．S [　　B　Ł

図 7のゴール到達問題ではエージ ェントはあらかじめ固定された初期座標からスタートし，固定した位置にあるゴール に到達すると正の報酬が得られ，エージェントが壁に衝突す ると罰が与えられる問題である．

エージェントの行動出力は前進，3 0度回転（左右），後退であり停止はない． また，回転した場合も後退した場合も前進した場合の半分の速度で移動は行われる． そのため，壁からある程度離れた段階で曲がり始めることが求められる．

エージェントのセンサ入力は自分と壁との距離を計測するセンサが角度（-60,-30,0,30,60)に5つ配置されているのに加え，ゴールの方向が与えられるセンサもあり，計6次元の連続値入力が与えられる．さらに，エージェント の状態遷移先は速度の5%の範囲で不確実になり，エージェントのセンサ

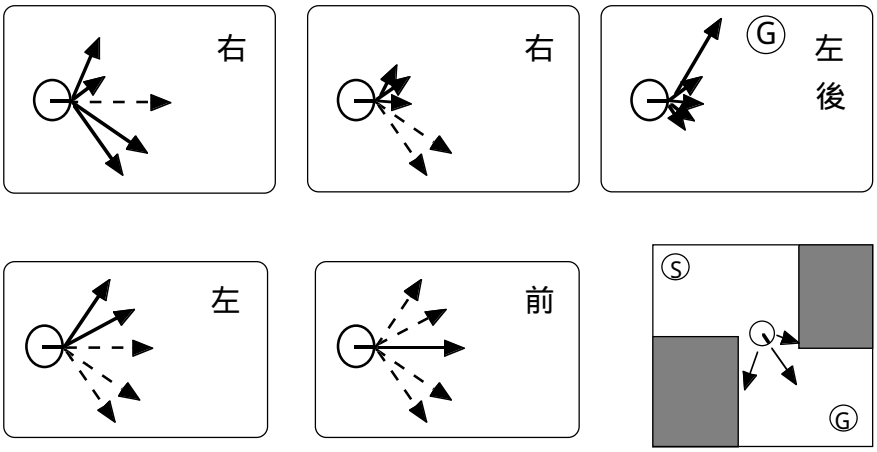


図8 ゴール到達問題分割典型例（矢印実線：政策に影響ある入力，矢印波線：政策に影響ない入力，右上の文字：政策）

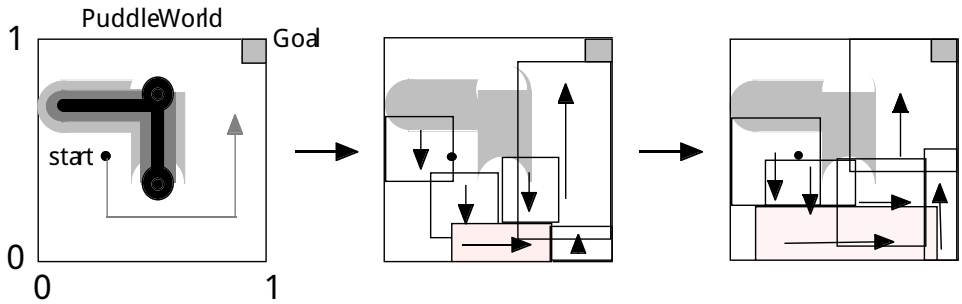


図9 PuddleWorld問題分割典型例

入力も最大可測距離の5％の範囲でノイズが加えられる設定とした．

この問題に対し，提案手法を適用した結果がグラフ4(5試行)である．このグラフ から提案手法による学習はある程度成功したといえる．また，内部状態数については 最大値を10に設定した．しかし，実際には学習初期には3 ~ 4 状態が生成され，300ゴール到達後には 6 ~ 8状態が存在するという結果であった．図8は分割と政策の典型例である。

6 次元の問題なのでたとえ各次元を2 等分ずつに分割したとしても6 4 状態になることを考慮すると，状態数に関しては提案手法は有効であるといえる． しかし、新たな状態を加えた直後に獲得報酬が改悪するという現象が観測された。

この問題に提案手法を適用した範囲では、提案手法は2次元問題の場合と比較して 状態の生成数が爆発するといったような問題は生じなかった。内部状態の学習が 不適切な場合、必要な観測点から内部状態が移動してしまい次々と追加が行なわれる といった現象が考えられる。提案手法ではそのような問題は生じなかったためと考えられる．

4.2.PuddleWorld Ł

図9のPuddleWorld問題では，エージェントはあらかじめ固定された初期状態からスタートし，ゴールに到達すると報酬が与えられ、水溜りに入ると深さに比例した罰が与えられる。 エージェントには自らの座標が連続値で与えられ、上下左右に進むという4種類の行動の選択が可能である。初期位置は図1 4 の黒点の位置としたため、エージェントは水溜りエリアを回避したのちにゴールしなければならない。 エージェントの行動遷移は先は1 ステップの歩幅の10%の幅でノイズを加えた。

この問題に、提案手法を適用したところ1 0 試行のうち5試行でほぼ最適な政策に収束した．図？に分割の典型例を示す。領域の中の矢印はその領域内の政策である。図9から，提案手法では適切な行動を学習した状態が境界線を広げるように学習が進んでいるようすが確認できる。図中の色のついた領域が 広がることにより水溜りのすぐ近くを通りすぎることが可能となり、 高い水準の報酬が学習可能となる。

5． Ａ ／ @

提案手法に関連のある研究、本研究との相違点について述べる。

・ 教師あり学習の分野では、特徴量空間の局所的部分をカバーするような 素子を用いた、Radial Basis Function Neural Networksがある。さらに、素子の発火の強度が弱い特徴量の領域に新たな中間層を追加する手法も 提案されている[Fritzke 92]。しかし、強化学習のような教師が存在しない状況でどのような学習を行なうべきかについての研究はあまりなされていない。

・ Fuzzy制御の分野では、提案手法と同様の適合度関数(メンバーシップ関数) を用いた制御器が数多く提案されており、Ｆｕｚｚｙ ＡＲＴ/ＡＲＴＭＡＰ [Carpenter 91]では、 教師が教えるルールが制御器に存在しない場合に新たなルールを加えている。Fuzzy ART/ARTMAP を強化学習問題に適用する際には、Actor-Critic Learning [Barto 83]と組み合わせることにより、強化学習問題を教師あり学習問題として解く手法が提案されている[Lin 96]。 環境モデルの学習が必要となる点が、提案手法との違いである。

・ 従来の強化学習の分野では、実験2のＰｕｄｄｌｅ World問題に対して様々な 手法を適用した結果が報告されている。線形ニューラルネットワーク、線形近似、 2 次近似等の近似システムと組み合わせると学習不可能との報告がある[Ｂｏｙａｎ ９５]。[Moore 95]は、観測空間を適応的にGrid分割する手法を提案し、適切な政策の学習を行なっている。しかし、Grid分割を用いているため比較的多くの分割が必要となっている。[Sutton 95]は、関数近似システムにCMAC[Albus 80]を用いた手法を提案し、 良好な結果を得ている。しかし，CMACをGridベースにした場合、適用可能な問題は比較的低次元の問題に限られると考えられる。また、CMAC のパラメータは 試行錯誤的に決定しなければならないという問題がある。

6． ¤ Ł

強化学習を高次元，連続値の問題に適用するため確率的な出力をする状態認識器を導入し，確率的傾斜法の更新式を利用して状態認識器のパラメータ及び行動選択器のパラメータを同時に学習させる手法を提案し，実験によって有効性を確認した．今後の課題は，内部状態表現の改良，及び他の問題への適用，他手法との比較である．

参考文献

[Watkins 92] Watkins, H. and Dayan, Technical Note Q-Learning, Machine Learning 8, pp. 279-292 (1992)

[Sutton 88] Sutton, Learning to Predict by the Methods of Temporal Differences, Machine Learning 3, pp. 9-44 (1988)[Lin 93] Lin, L. J., Scaling Up Reinforcement Learning for Robot Control,Proceedings of the 10th International Conference on Machine Learning,pp. 182-189 (1993)

[Tan 91] Tan, M., Cost-Sensitive Reinforcement Learning for Adaptive Classification and Control, Proceedings of the 9th NCAI, pp. 774-780 (1991)

[Tesauro 92] Tesauro G,Temporal Difference Learning of Backgammon Strategy,9th ICML, pp. 451-457 (1992)

[Satinder 97] Satinder Singh, Dimitri Bertsekas, Reinforcement Learning for Dynamic Channel Allocation in Cellular Telephone Systems, NIPS9, pp.974-980(1997)

[Robert 96] Robert H.Crites, Improving Elevator Performance Using Reinforcement Learning , NIPS 8(1996)

[Mahadevan 97] Mahadevan, S., Marchalleck, N., Das, T.K., Gosavi,A., Self-Improving Factory Simulation using Continuous-time Average-Reward Reinforcement Learning, Proccedings of the 14th ICML pp.202-210(1997).

[Boyan 95] Justin A.Boyan, Generalization in Reinforcement Learning Safely Approximationg the Value Function, NIPS 7, pp.369-376(1995)

[Leemon 95] Leemon Baird, Residual Algorithms: Reinforcement Learning with Function Approximation, 12th ICML, pp.30-37 (1995)

[Williams 92] Williams, R.j, Simple Statistical Gradient Following Algorithms for Connectionist Reinforcement Learning, Machine Learing 8,pp.229-256(1992)

[Kimura 95] Kimura, H., Yamamura, M. and Kobayashi, S., Reinforcement Learning by Stochastic Hill Climbing on Discounted Reward,12th ICML pp.295-303,(1995)

[Singh 94] Singh, S.P.Jakkola, T and Jordan, M.I.:Learning without State-Estimation in Partially Observable Markovian Decision Processes, 11th ICML, pp.284-292(1994)

[Moore et al. 95] Moore A. W. Atkeson, C. G.: The Parti-game Algorithm for Variable Resolution Reinforcement Learning in Multidimensional State-spaces, Machine Learning 21, pp. 199-233 (1995).

[Fritzke 94] B.Fritzke : Growing Cell Structures, Neural Networks, 7, 9, pp.1441-1460(1994)

[Carpenter 91] G.A.Carpenter, S. Grossberg, and D. B. Rosen : Fuzzy ART: Fast stable learning and categorization of analig patterns by an adaptive resonance system, Neural Networks, vol. 4, pp. 759-771, 1991

[Barto 83] A. G. Barto, R. S. Sutton, and C. W. Anderson : Neuronlike adaptive elements that can solve difficult learning control problem, IEEE Trans. Syst., Man, Cyber., vol. SMC-13, no. 5, pp. 834-847, 1983.

[Lin 96] Cheng-Jian Lin and Chin-Teng Lin : Reinforcement Learning for An ART-Based Fuzzy Adaptive Learning Control Network, IEEE TRANSACTIONS ON NEURAL NETWORKS, vol. 7, no. 3, pp. 709-731, 1996

[Sutton 95] Richard S. Sutton : Generalization in Reinforcement Learning : Successful Examples Using Sparse Coarse Coding, NIPS 8, pp. 1038-1044

[Albus 80] JAMES S.ALBUS : Mechanisms of Planning and Problem Solving in the Brain, Mathematical Biosciences, 45, pp. 247-293