修士論文

アクター．クリティックにおける禁止行動規則に着目した転移学習

平成21年度修了
三重大学大学院工学研究科
博士前期課程　電気電子工学専攻

高野　敏明

目 次

第1章　はじめに

近年，ロボットは産業用だけでなく，医療用や家庭用として開発されるようになっている [1]．これはロボットが工場などの整備された環境下だけでなく，時々刻々と変化する環境下においても与えられたタスクをこなせるようになってきていることをあらわしており，ロボットの活躍する範囲が広がってきている証拠である．将来的には，人間に代わり危険な作業を行ったり，ロボットが人間と自然な会話をすることなどが期待されている．このようなタスクには，人間が予測不可能な事態が発生する可能性が十分にある．このような環境下でロボットがタスクを行うには，ロボット自身が不測の事態に対して柔軟な行動をとることが要求される．そのため，人間がロボットの行動規則をプログラムするだけでなく，ロボット自身が学習を行い，その環境下で行うタスクに対する行動規則を獲得する必要がある．

ロボット自身がある環境下で学習を行うことにより，タスクを達成するための手法として強化学習 [1] が研究されている．強化学習は，工場のような整備された環境はもちろん，宇宙空間などの予測困難な環境下での応用が期待されている．しかし，強化学習は与えられたタスクに対して適切な行動規則を獲得するまでの学習回数が多いという問題が指摘されている [2]．また，タスクにおける行動目標が同一でも，タスクを行う環境が異なる場合，学習した行動規則により，タスクが達成できないことがある．このような場合，再びタスクを学習する必要があり，この学習にも多くの学習回数が必要となる．以上の理由により，現段階において強化学習を実用化が困難である [3]．強化学習を実用化するには，このような問題を解決する必要がある．

強化学習を実用的なものにするため，学習回数を削減する研究が多くなされている．学習回数の削減する従来研究の例として，学習パラメータの最適化 [5]，モデルベース強化学習 [6]，転移学習 [7] などが挙げられる．学習パラメータの最適化やモデルベース強化学習は，単位学習回数あたりの学習達成度を多くすることで，学習回数の削減を図る手法である．一方，転移学習は，学習開始前や学習途

中から学習達成度をある程度引き上げることで，学習回数の削減を図る手法である．本研究では，この転移学習に着目する．特に，タスクごとに異なる環境下での適切な行動規則を少ない学習回数で獲得するために，過去に獲得した知識を現在のタスクの学習に利用することで学習回数を削減する方法について検討する．

強化学習の代表的な手法として，Profit Sharing，Q学習，アクタークリティックなどが挙げられる．強化学習を実用的なものにすることを考えると，Profit Sharingは学習は早いが，収束性にかける．また，Q学習では，連続値行動のような，可能な行動の数が無限大の場合，状態価値を学習する方法では，1つの行動を選び出すために無限集合の中を探索することになり，莫大な計算を必要とする．アクター・クリティックは，その収束性が証明されており [10]，行動を陽に表現しているため，連続値行動においても莫大な計算を必要としない [1]．以上の理由により，本研究では，強化学習の一手法であるアクター・クリティックを対象に，その学習回数を削減する手法について検討する．

本論文の構成は以下のとおりである．2章では，強化学習の枠組みと本研究で対象とするアクター・クリティックに関する説明を行う．3章では，強化学習の問題点を解消するための従来研究について紹介し，各研究の問題点を述べる．また，転移学習についてもこの章で述べる．4章では，3章で述べた問題点を解消するための方法を提案する．5章では，4章で提案した方法によって学習回数の削減が行えるかを検証する．最後に6章で，本研究をまとめる．

第2章　強化学習

本章では，本論文で対象とする強化学習と，強化学習の代表的な手法の一つであるアクタークリティックについて説明する．

2.1　強化学習

強化学習は，状態，行動，そして報酬に関してエージェントとその環境との間の相互作用を定義している形式的な枠組みである [1]．ここで，強化学習において，一般に，エージェントとは学習と意思決定を行うものをさす．また，エージェントが相互作用を行う対象を環境とよぶ．強化学習においてエージェントは環境と以下のやりとりを行う (図 2.1)．

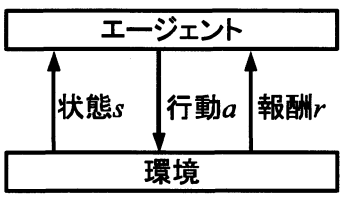

図 2.1: 強化学習の枠組み

1. エージェントは環境から状態 s を観測する．
2. エージェントは観測した状態 s に基づいて行動 a を選択，実行する．
3. エージェントは，その行動 a の結果，行動 a に対する価値に応じた報酬 r を環境から与えられる．
4. 1〜3 を繰り返す．

エージェントは環境との相互作用の中で，報酬を手掛かりに，与えられたタスクのすべての状態における行動規則を獲得する．ここで，タスクとは，ある環境下において，あらかじめ定められた所定の状態（以下，ゴールとよぶ）に到達することである．また，行動規則とは，ある状態においてどの行動をすればよいかを表した規則である．一般に，エージェントが与えられたタスクにおいてゴールに到達した際には正の報酬，ある状態において誤った行動を行った際には負の報酬が与えられる．強化学習のアルゴリズムは，一行動あたりにうけとる報酬の期待値を大きくするように定式化されている．

強化学習の環境モデルはしばしば，マルコフ決定過程 (Markov Decision Process, MDP) によって定式化される．MDP とは，マルコフ性 (次状態 s' への遷移が現状態 s と行動 a にのみ依存し，それ以前の状態や行動には関係しない性質) を満たした環境のことである．MDP にもさまざまなモデルがある．代表的な MDP をいかに示す．

- 単純マルコフ決定過程 (単純 MDP)
- セミマルコフ決定過程 (Semi-Markov Decision Process, SMDP)
- 部分観測マルコフ決定過程 (Partially Observable Markov Decision Process, POMDP)

これらについて説明する．単純MDPは，上で述べたMDPそのものである．SMDP[8]は，現状態 s と次状態 s' の状態観測の時間間隔が任意の MDP である．POMDP[9]は，エージェントの状態観測が不確実性や不完全性を持つ MDP である．このように，より実世界に近い環境モデルを構築するために，さまざまな MDP が提案されている．

本研究では, 問題の定式化を簡単にするため, エージェントがとりうる状態数や選択可能な行動数が有限である離散時間での単純 MDP を対象とする．

2.2 アクター・クリティック

1章で述べたように，アクタークリティックは収束性が証明されている [10]．また，行動を陽に表現している [1] ため，連続値行動に対しても有効である．これらをふまえて，本研究では，強化学習の一手法としてアクター・クリティックを用いる．アクター・クリティックにおいて，エージェントはアクター (行動器) とクリ

ティック（評価器）により構成されている (図 2.2)．アクターは，環境から状態を観測し，その状態に合わせて行動を選択する．クリティックはこの行動の結果として得られる報酬から，TD 誤差 δ(Temporal Difference) を算出し，この TD 誤差により，アクターの選択した行動を評価・更新する．アクター・クリティックの学習において，エージェントがとりうるすべての状態集合を S，選択可能な行動集合を A とし，状態 $s \in S$ における行動 $a \in A$ の価値を行動優先度 $p(s,a)$ とよぶ数値で表し，ある状態の推定の価値を状態価値 $V(s)$ とよぶ数値で表す．アクター・クリティックでは，これらのパラメータの修正を繰り返すことで学習する．行動優先度の修正は，ある状態 s において適切な行動優先度 $p(s,a)$ が最大になるように更新され，状態価値の修正は，ゴールに到達するために通過しなければいけない状態でゴールに近い状態ほど大きくなるように更新される．アクター・クリティックにおける更新式を以下に示す．

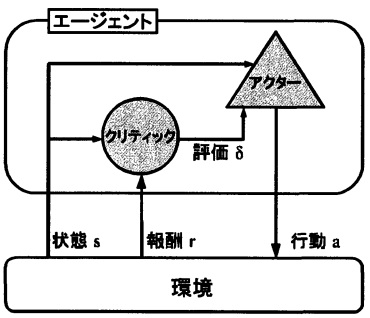

図 2.2: アクター・クリティックの枠組み

$$\delta = r + \gamma V(s') - V(s) \tag{2.1}$$

$$V(s) \leftarrow V(s) + \alpha\delta \tag{2.2}$$

$$p(s,a) \leftarrow p(s,a) + \beta\delta \tag{2.3}$$

ここで，式 2.2，式 2.3 において左矢印 ($\leftarrow$) は左辺の変数に右辺の値を代入する操作を表している．また，δ は TD 誤差を表し，r は状態 s において行動 a を行った結果，得られる報酬を表す．$V(s)$ は状態 s における状態価値，s' は状態 s において行動を a を行った結果，遷移した先の状態，$p(s,a)$ は状態 s において，行動 a をどれほど優先して行うべきかを表している．γ，α，β はそれぞれ，割引率，学習率，ステップサイズパラメータとよばれるあらかじめ与えられたパラメータ (以

降，学習パラメータとよぶ) である ($0 \leq \gamma \leq 1$，$0 \leq \alpha \leq 1$，$0 < \beta$)．割引率 γ は将来獲得予定の報酬を現時点でどれだけ重要視するかの割合を表したものである [11]．また，学習率 α は現時点での $V(s)$ と，報酬や遷移した先の状態から得られる結果とのバランスを表したものである．同様に，ステップサイズパラメータ β は $p(s,a)$ において報酬や遷移した先から得られる結果とのバランスを表したものである．アクター・クリティックでは，α，β を十分小さい値にとることで，これらの値にかかわらず 1 つの解に収束する [1][10]．

これらの更新式により，行動優先度と状態価値の更新を繰り返すことで，各状態における適切な行動の行動優先度が最大になるように修正する．ここで，適切な行動とは，アクター・クリティックではゴールに向かう行動のことを指す．最終的に，各状態における最大の行動優先度を持つ行動が，与えられたタスクにおける各状態での行動規則となる．

アクター・クリティックは，学習パラメータが適切に設定されていれば，無限回の学習を行うことで学習が収束することが証明されている [10]．また，行動が陽に表現されている [1]．そのため，強化学習のさまざまな研究においてアクター・クリティックが用いられている．問題点としては，強化学習は試行錯誤により，適切な行動を学習するため，学習が完了するまでには多数回の学習が必要になることが広く知られている [2]．

2.3 行動選択法

アクター・クリティックでは，学習パラメータを適切に設定し，無限回の学習を行うことで学習が収束することは前節で述べた．しかし，実世界で無限回の学習を行うことは無限の時間を費やすことになり，非現実的である．そこで，学習の収束を早める行動選択法がいくつか提案されている [1]．代表的な行動選択手法を以下に示す．

- greedy 法
- ϵ−greedy 法
- ボルツマン選択

各行動選択手法について説明する．greedy 法とは，現状態 s において最大の行動優先度 $p(s,a)$ を持つ行動 a を決定的に選択する手法である．アクター・クリティック

では，各状態における適切な行動の$p(s,a)$の値を最大にするように学習するため，greedy法は最も単純な行動選択手法である．ϵ−greedy法は，確率$1-\epsilon$でgreedy法によって行動を選択し，確率ϵで行動優先度によらずランダムに行動を選択する手法である．この手法は確率ϵでgreedy法とは異なった行動を選択できる可能性があり，この場合，greedy法とは異なった状態価値や行動優先度を更新できるという利点がある．なお，$\epsilon=0$のとき，ϵ−greedy法はgreedy法と等価である．最後に，ボルツマン選択とは，状態sにおける行動行動優先度$p(s,a)$の比によって確率的に行動を選択する手法である．状態sにおいて,行動aを選択する確率$p(a|s)$は以下の式によって求める．

$$p(a|s)=\frac{\exp(p(s,a)/T)}{\Sigma_b \exp(p(s,b)/T)} \tag{2.4}$$

この式より，大きな行動優先度を持つ行動ほど選択される確率が高く，小さな行動優先度を持つ行動ほど選択される確率は低くなる．なお，Tは温度パラメータとよばれ,$T\to 0$のとき，greedy法と等価になる[1].

いずれの行動選択手法においても，学習の進行を促すようなパラメータの設定が難しい．また，現状ではどの行動選択手法をとるべきかという明確な指針は現状では示されていない[11].

第3章　強化学習における学習の高速化

強化学習により，タスクに対する行動規則を学習によって自律的に獲得できることは2章で述べた．その一方，学習回数が多いなどの問題点があり，現段階では実用化が困難である [3]．そのため，実用化に向けてこの問題点を解消するための研究が多くなされている [5][6][13][14][15].

本章では，強化学習の問題点を解消するための従来法をいくつか紹介し, 各従来法の問題点をふまえたうえで，強化学習の高速化の検討を行う．

3.1　従来研究

学習の高速化に関する主な従来研究を以下に紹介する．学習の高速化に関する従来研究のねらいは次の2つに分類できる．

1. 現在行っているタスクにおける単位学習回数あたりの学習達成度の上昇率を上げる．

2. 現在行っているタスクの学習開始時や学習途中で学習達成度を引き上げる．

ここで，学習達成度とは, 適切な行動をとることができる状態の全状態に対する割合のことを指す．学習達成度が0% に近いほど，適切な行動をとることができる状態が少なく，学習達成度が100% に近いほど，適切な行動をとることができる状態が多いことになる．前者は，学習の仕方を工夫することで，学習達成度の上昇率を上げるものである．一方，後者は学習とは別の方法で学習達成度を引き上げるものである．それぞれのねらいのもとでの従来研究を簡単に説明する．

まず，前者をねらいとする従来研究として，パラメータ設定の最適化 [5] とモデルベース強化学習 [6] の2つを紹介する．パラメータ設定の最適化は，強化学習の学習パラメータである，学習率α や割引率γ などを遺伝的アルゴリズムによって

最適な数値となるように設定・調整する手法である．これは，学習パラメータを適切に設定することで，学習によって変化するパラメータを適切に更新され，学習回数を削減する狙いがある．

モデルベース強化学習は，エージェントの内部に環境モデルを作り，実際の学習とモデル内の学習の2種類の学習を行うことで，学習によって変化するパラメータを更新する手法である．これは，いわば人間がイメージトレーニングをしたうえで動作を行うのと同様である．実際の行動は1回でも複数回のパラメータの更新を行うことになる．これにより学習回数の削減をねらう．ここで，行動1回とは，エージェントが行動を選択・実行した回数を表す．また，学習1回は，エージェントが状態を観測する，行動を選択・実行する，その結果として報酬をうけとる，各パラメータを更新する，までを意味する．

次に後者をねらいとする従来研究の転移学習 [7] を紹介する．転移学習は，過去に獲得した知識を，これから学習するタスクに利用する手法である．これは，現在行っているタスクの学習開始時や学習途中から過去に経験した知識を利用することで，適切な行動をとることができる状態を多くし，学習回数を削減するねらいがある．

われわれ人間は，現在直面している問題を学習するときには，0から学習するのではなく，過去に学習したタスクで獲得した知識を利用している．そこで，強化学習においても，ある環境で現在行おうとしているタスクにおける学習回数を削減するためには，0から学習を開始するのではなく，過去のタスクで獲得した知識を利用することが自然である．しかし，冒頭で述べたように，強化学習では，同一の行動目標であっても異なる環境下ではうまく働かない．このうまく働かない要因を取り除くことができれば，学習回数の削減につながることはいうまでもない．

以上により，本研究では，転移学習に着目する．次節で，転移学習の従来研究について説明する．

3.2 転移学習

転移学習とは，過去に学習したタスク (以降，過去のタスクとよぶ) により獲得した知識 (以降，過去の知識とよぶ) を何らかの手法により，現在行おうとしているタスク (以降，現在のタスクとよぶ) に用いることである．転移学習として関係の深い研究として，帰納転移 [19]，マルチタスク学習 [12]，知識の再利用 [13] など

といった研究がある．これらの手法の基本的な手順を以下に示す．

手順1.　過去の知識の蓄積．

手順2.　利用する（転移させる）知識の選択．

手順3.　選択した知識を転移．

手順4.　現在のタスクを学習．

転移学習を行ううえで最も重要なことは，手順2と手順3である．これらを効果的に行うことで，転移学習により学習回数を削減できる．手順2の転移させる知識をどのように決めるかは，転移直後の学習達成度を引き上げるために重要なことである．そのためには，現在のタスクに対する適切な行動と同じ行動を行う状態の割合が高いものを選ぶべきである．学習達成度を引き上げる知識を選んだとしても，その知識の中に利用してはならない行動規則がいくつかは存在する．この利用してはならない行動規則を利用した場合，それらの行動規則の修正に学習回数を要してしまい，逆に学習回数を増加させてしまう可能性がある．そのため，手順3では，利用してはならない行動規則を判別し，現在のタスクに利用しないように過去の知識を分別したうえで利用する，あるいは，少ない学習回数で過去の知識を修正できるように，過去の知識を加工して利用することを行う．これらを解消することで，効果的に転移学習を行えると考えられる．

ここで，強化学習における知識とはなにかについて考える．知識とは，ある事柄に対する明確な意識と判断である．これを強化学習で考えると，与えられたタスクのエージェントが認識できるすべての状態における適切な行動が明確に判断できることである．つまり，ある状態における適切な行動の集合が知識である．アクター・クリティックでは，このある状態における行動を左右するパラメータとして行動優先度がある．また，行動優先度を更新するためには，状態価値が必須であり，状態価値なくしては学習ができないことから，本論文では，状態価値も知識であるとみなす．つまり，アクタークリティックでは行動優先度と状態価値が知識である．この知識を転移学習によって利用する方法はさまざま考えられる．以下で，転移学習によって学習回数を削減する従来の研究を紹介し，各手法による効果と課題を検討する．

- PRQ-Learning(Policy Reuse in Q-Learning)[14]
  この文献では，実際に過去の知識を用いて現在のタスクで行動をさせ，その

結果から過去の知識として最も適した過去の知識を選び，過去の知識で現在のタスクを行動させ，その裏で学習を行う手法について提案している．このとき，エージェントの行動は，確率 φ で過去の知識に基づいた行動を行い，確率 $1-\varphi$ で学習中の知識を用いて ϵ−greedy 法で行動を行う．この方法であれば，過去の知識の中で最も学習達成度の高い知識を選ぶことができる．しかし，この方法は過去の知識で現在のタスクが達成できなければ過去の知識を選ぶことができない．つまり，環境がほぼ同一で，適切な行動がかなり一致していることが前提となってしまう．

- 強化学習結果の再構築への概念学習の適用 [15]
  この文献では，学習開始前に概念学習を用いて過去のタスクと現在のタスクを比べ，過去の知識が利用できるかを判断する手法を提案している．この手法では，学習開始前に現在のタスクについての背景知識が得られていることが前提である．しかし，もし，過去の知識の中に利用できる知識が存在しない場合は，学習回数が増加するという問題は，この文献内でも指摘されている．

以上から，本研究では，転移学習の学習回数削減の効果を高める方法について検討する．従来手法で問題となった，適切な行動が多少異なっていても過去の知識を選ぶことができ，かつ，過去の知識に現在のタスクに利用すべき知識が存在しない場合でも，現在のタスクの学習回数が増加しない方法についてアクター・クリティックを用いて検討する．

第4章　提案法

前章では，学習の高速化に関する従来研究として，特に転移学習について述べた．本章では，転移学習における各手法の課題をふまえたうえで，効率的に転移させる知識を見つける手法とアクター・クリティックの特徴に着目した知識の転移方法について提案する．

4.1　効果的な転移学習

転移学習を行う場合，以下のような疑問がある

- どのような過去の知識を蓄積するか．
- 知識の転移をさせるべきか，あるいはどの知識を転移させるべきか．
- 転移させるならどのように知識を転移させるべきか．

一点目は，3.2節において示した手順1についての疑問，二点目は3.2節において示した手順2についての疑問，三点目は3.2節において示した手順3についての疑問を表している．

一点目は，どのような知識を過去の知識としてデータベースに残すかである．データベースにすべての学習結果を残せば，さまざまな知識を残すことができる．しかし，知識を転移させようとした場合，現在のタスクの学習回数を少なくするような知識の発見までに時間がかかってしまう．そのため，データベースに残す知識は取捨選択する必要がある．

二点目は，知識の転移をさせるか否か，あるいはデータベースのどの知識を転移させるべきかである．知識を転移させることで，学習回数を削減するためには，転移させる知識が現在のタスクにおいて獲得すべき知識と類似している必要がある．もしこれが，類似していない場合，誤った知識を修正するために，かえって学習回数が増加してしまう．また，この判別に学習回数を費やすことも，全体と

して学習回数の増加につながってしまう．そのため，少ない学習回数で転移させるのに有効な知識を分別する必要がある．

三点目は，選ばれた知識の転移方法である．理想としては，選ばれた知識を転移させることによって，即刻，タスクを達成できることが理想である．しかし，このような状況はまれで，ほとんどの場合，選ばれた知識を，現在のタスクに合うように修正しなければならない．そのため，この修正が容易であれば学習回数の削減につながる．

これらの疑問を解消することで，転移学習を効果的に行うことができる．ただし，本論文では，すでにデータベースに知識が蓄積されているものとして議論を進める．

4.2 転移させる知識の選択

4.2.1 行動優先度による判定

2.2節で述べたようにアクタークリティックは，行動規則を行動優先度により表している．そこで，学習中の行動優先度により獲得すべき行動規則を予測することを考える．

行動優先度は，状態価値の大きな状態へ向かうように更新される (式2.3)．しかし，学習開始直後においては，最も大きな状態価値をもつ状態がゴールであるとは限らない．そのため，学習中の行動優先度はゴールではない状態へ向かう可能性がある．つまり，学習中の行動優先度をみても，ある状態において最大の行動が，選択されるべき行動であると判断することは不可能であると考えられる．これを，簡単な実験（5.1節と同等の実験）を行い，ある状態の行動優先度が学習によりどのように変化するのかを観測した．その結果を，表4.1に示す．

表4.1において，最適な行動は行動3であった．表4.1より，学習途中の各時点において，必ずしも行動3に対する行動優先度が，全行動中で最大となっていない．これにより，学習途中の行動優先度から，現在のタスクにおいて最適な行動が何であるかを推測することは困難であることが分かる．

しかし，現在のタスクにおいて使用してはならない行動規則を推測することは可能であると考えられる．ここで，使用してはならない行動規則とは，タスクの失敗が即刻確定するような行動規則（禁止行動規則）のことを指す．これは，タスク達成時に低い行動優先度をもつ行動（表4.1の行動1と行動4）について，学

表 4.1: 学習中における行動優先度の変化

	行動1	行動2	行動3	行動4
学習序盤 (20[episode])	-2.7	0.4	1.3	-3.0
学習中盤 (50[episode])	-2.7	0.8	-2.6	-3.0
学習終盤 (70[episode])	-5.0	-0.9	-5.2	-4.6
タスク達成 (90[episode])	-7.1	-4.6	4.2	-6.7

習中に一定して低い行動優先度を持っているためである．ただし，学習中盤から終盤にかけて，行動3も低い行動優先度を持っているため，行動優先度のみでは推測できない．

この傾向は他の状態・他の学習過程でも多くみられたことから，本研究では，禁止行動規則を抽出し，それを用いて転移させるべき知識を選択することを試みる．

4.2.2 禁止行動規則と転移する知識の選択

前項での議論をふまえて，禁止行動規則を抽出する方法，および，それを用いた知識の選択方法について検討する．

多くの場合，タスクの失敗を引き起こす行動に対しては，大きな負の報酬が与えられる．そのため，禁止行動規則は，学習過程で与えられる報酬を観測することで容易に検知できる．同様に，過去のタスクに対しても学習中に報酬を観測することで禁止行動規則を得ることができる．禁止行動規則は，状態と行動の組として保持される．

以上により抽出した禁止行動規則を，現在のタスクと過去のタスクとで比較することで，転移させるべき知識の選択ができる．具体的には，抽出した禁止行動規則を集め，各状態ごとに禁止行動のリストを作成する．現在のタスクと過去のタスクとでこの禁止行動のリストが一致している状態 (等価状態) が多くなるほど，転移学習の効果が高いと考え，そのような知識を転移させる知識として選択する．ただし，等価状態が全状態に占める割合 (等価状態率) がしきい値θ(有効転移率) よりも下回る場合は，知識の転移を行わない．

4.3 知識の転移方法

4.3.1 禁止行動規則による転移させる知識の選択方法の性質

転移させる知識の選択方法の性質について議論する．禁止行動規則を利用した知識の選択法により，学習中にある程度の行動規則を知ることが可能となると考えられる．しかし，この選択法を用いても，エージェントが学習中であるため，すべての状態を経験していないことが容易に考えられる．とくに，学習中盤から経験することになるであろう状態は過去の知識と異なる可能性があることが容易に考えられる．そこで，このような事態をふまえ，過去の知識を転移させる必要がある．詳細な知識の転移方法については次項以降で説明する．

4.3.2 行動優先度の転移方法

行動優先度は，エージェントの行動を決定するためのパラメータである．もし，「間違った行動優先度」を転移させた場合，その間違いを修正するために通常より多くの学習が必要となる．そのため，行動優先度を転移させる際には，「間違った行動優先度」を転移させるべきではない．ここでいう，「間違った行動優先度」とは，ある行動優先度において，過去の知識では最も選ばれる行動であるのに対し，現在の知識では，タスクの失敗が即刻確定する場合，あるいは，過去の行動ではタスクの失敗が即刻確定するが，現在のタスクでは，選ばれるべき行動の場合，その行動優先度は「間違った行動優先度」であると定義する．「間違った行動優先度」を転移させないことでタスクの学習回数を増加させる知識を利用しないと考えられる．その結果として，現在のタスクにおいて学習回数を削減することが期待できる．

現在のタスクと過去のタスクが等価状態にあるか否かによって，「間違った行動優先度」を判別できると考える．判別した結果，等価状態の場合，その行動優先度を転移させる．このとき，過去の知識だけでなく，現在のタスクの学習によって得られた知識を反映させるため，現在の行動優先度に過去の行動優先度を足し合わせる．

知識の転移させる場合，現在のタスクの行動優先度を $p(s,a)$，過去のタスクの行動優先度を $p_b(s,a)$ としたとき，s が等価状態の場合に現在の行動優先度を式 4.1

により更新し，そうでない場合には行動優先度は更新しないこととした．

$$p(s,a) \leftarrow p(s,a) + p_b(s,a) \qquad \{\forall a \in A\} \tag{4.1}$$

4.3.3 状態価値の転移方法

状態価値は，エージェントの行動を評価するためのパラメータである．もし，間違った状態価値を与えると，エージェントの行動を間違った行動へと導いてしまい，結果として，転移学習を行わない場合と比べ多くの学習回数を要してしまう．しかし，状態価値は行動優先度ほどエージェントの行動に直接影響を及ぼすものではない．そのため，行動優先度を利用するより，低いリスクで，エージェントの行動を導くことができる．

行動優先度の転移方法では，「間違った行動優先度」を転移させないように禁止行動規則を利用する方法を提案したが，状態価値の転移では，禁止行動規則からは，知ることのできない部分において過去のタスクの行動規則へと導くように転移させることを考える．4.2.1節で述べたように，状態価値の大きな方へ，行動優先度は導かれる．そのため，過去のタスクの行動規則へと導くためには，値の大きな状態価値を転移させることにより実現できると考えられる．つまり，正の値を持つ状態価値を転移させる．このとき，過去の知識に現在の知識を加えるだけでなく，多少は間違った状態価値が含まれていることを考え，間違った状態価値による行動優先度のミスリードを低減させるため，過去と現在の状態価値を平均化する．

知識を転移させる場合，現在のタスクの状態価値を$V(s)$，過去のタスクの状態価値を$V_b(s)$とすると，$V_b(s) > 0$の場合に式4.2に従い，現在の状態価値を更新し，そうでない場合には状態価値を転移しないこととした．

$$V(s) \leftarrow \{V(s) + V_b(s)\}/2 \qquad \{V_b(s) > 0\} \tag{4.2}$$

4.4 提案法まとめ

4章で述べた提案法の手続きは以下のとおりである (図4.2)．

手順 (i) では，過去に学習したタスクの学習結果をデータベースに記録する．このとき，学習したタスクの行動優先度$p_b(s,a)$，状態価値$V_b(s)$，禁止行動規則$f_b(s)$，

タスクの詳細な設定などを記録する．これらをいくつかの過去のタスクについて行うことでデータベース D が作成される．

手順 (ii) では，これから学習したい現在のタスクについて，初期化を行う．初期化として，行動優先度 $p(s,a)$，状態価値 $V(s)$，禁止行動規則 $f(s)$ を初期化する．また，学習パラメータである割引率 γ，学習率 α，ステップサイズパラメータ β，しきい値 θ を設定する．

手順 (iii) は，提案法のアルゴリズムを表している．(a) から (d) までが一般的なアクター・クリティックのアルゴリズムである（ただし，禁止行動規則の記録は除く）．提案法は，(c) において報酬を監視し，この値がタスクの失敗を意味するような負の値であれば，(b) で行った行動を禁止行動と認識し，$f(s)$ に記録する．また，(e) では，各過去のタスクと現在のタスクの等価状態率を算出し，有効転移率 (しきい値 θ) を超えた過去のタスクの中で，等価状態率が最も大きなものを転移させる知識として選択する．ただし，選ばれたものが，最後に転移した知識と同一である場合には，転移を行わない．

(i) 過去のタスクを学習する．データベースDの作成.
行動優先度 $p_b(s,a)$，状態価値 $V_b(s)$，禁止行動規則 $f_b(s)$
タスクの詳細な設定などを記録する.

(ii) 初期化

(a) 現在のタスクを用意する

(b) 行動優先度 $p(s,a)$，状態価値 $V(s)$，禁止行動規則 $f(s)$ の初期化

(c) 学習パラメータ γ，α，β，θ の設定

(iii) For i = 1 to "K"または"タスク達成" do

(a) 状態 s を観測する.

(b) 行動 a を決定・実行する.

(c) 報酬をうけとる.
a が禁止行動であれば，$f(s)$ に記録する.

(d) $p(s,a)$，$V(s)$ を更新する.

(e) 有効転移率を測り，転移させるべき知識がみつかれば，
その知識を転移させる.
禁止行動規則が一致している各状態 s において，
$$p(s,a) \leftarrow p(s,a) + p_b(s,a) \qquad \{\forall a \in A\}$$
転移させる状態価値が正のものについて
$$V(s) \leftarrow \{V(s) + V_b(s)\}/2 \qquad \{V_b(s) > 0\}$$

図 4.2: 提案法の流れ

第5章　検証実験

この章では，転移する知識の選択方法を用いたうえで，行動優先度と状態価値を前章で説明したように現在のタスクに利用する．このとき，全体の学習回数を削減できるかを，簡単な実験により検証する．

5.1　実験条件

実験は，現在の位置を観測し，上下左右のいずれかの行動をするエージェントに，簡単な迷路のスタートからゴールまで走破させるタスクを学習させることを行った．図5.1にエージェントに実験で使用した迷路を示す．迷路は7×7の格子

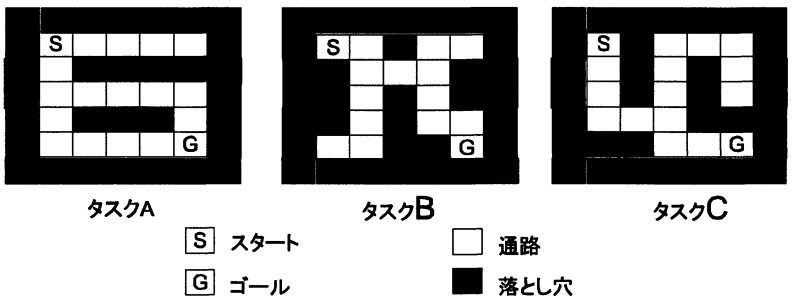

図 5.1: 現在のタスク

から成り，白マスが通過してもよい通路，黒マスが侵入できない壁を意味している．図中，Sはスタート，Gはゴールを意味する．

実験では，過去のタスクとして，図5.2に示すような迷路を用意し，それぞれのタスクを十分に学習させた結果(行動優先度，状態価値，禁止行動規則など)をデータベースに保持している．

なお，実験の条件は以下の通りに定めた．報酬rは，壁にぶつかる行動に対して，$r = -50$，ゴールにたどり着いたら，$r = 100$，上下左右のいずれかの行動を100ステップ行うごとに$r = -25$を与えた．また，アクタークリティックの学習パラメータは，割引率$\gamma = 0.95$，学習率$\alpha = 0.05$，ステップサイズパラメータ

$\beta = 0.05$ とし，学習中の行動選択には ϵ−greedy 法を使用し，$\epsilon = 0.05$ とした．また，転移させる知識の選択方法では，データベースにある，すべての過去のタスクに対して等価状態率を算出し，有効転移率 θ 以上となる過去のタスクの学習結果を転移させる知識とした．その有効転移率は $\theta = 0.3$ とし実験を行った．これらの実験を 1000[episode] 行い，これを 30[trial] 繰り返した．なお，学習終了条件は，10[episode] 連続でゴールできるか，1000[episode] 学習を繰り返すかである．また 1[trial] とは，学習開始から学習終了までを表す．実験結果は 5.3 節で示す．実験では，禁止行動規則により転移する過去の知識を決定した後，行動優先度，状態価値のそれぞれのパラメータについて，転移学習を行わない，そのまま転移，提案手法による転移，の 3 手法で転移を行う．全体としては，以下に示す 9 通りの組み合わせを比較することで検証を行う．

- 転移学習を行わない (Normal)
- そのまま転移 (Simple)
- 提案手法 (Proposed)
- $p(s,a)$ はそのまま，$V(s)$ は提案手法で転移 ($\mathrm{p}_{all} - \mathrm{V}_{pro}$)
- $p(s,a)$ は提案手法，$V(s)$ はそのまま転移 ($\mathrm{p}_{pro} - \mathrm{V}_{all}$)
- $p(s,a)$ だけを提案手法で転移 (p_{pro})
- $V(s)$ だけを提案手法で転移 (V_{pro})
- $p(s,a)$ だけをそのまま転移 (p_{all})
- $V(s)$ だけをそのまま転移 (V_{all})

Normal は，転移学習を行うことなく通常のアクタークリティックで学習した場合，Simple は，本論文で提案した転移する知識の選択方法を用いて，転移する知識を選び，過去タスクの行動優先度と状態価値を現在のタスクの行動優先度と状態価値に足し合わせた方法である．Proposed は，前章で説明したように，行動優先度と状態価値を利用した方法である．p と V の添字について説明する．添字 pro は p と V のそれぞれを前章で説明したように利用する方法を用いることを示している．また，添字 all は p と V のそれぞれをそのまま転移させる方法を用いることを示している．

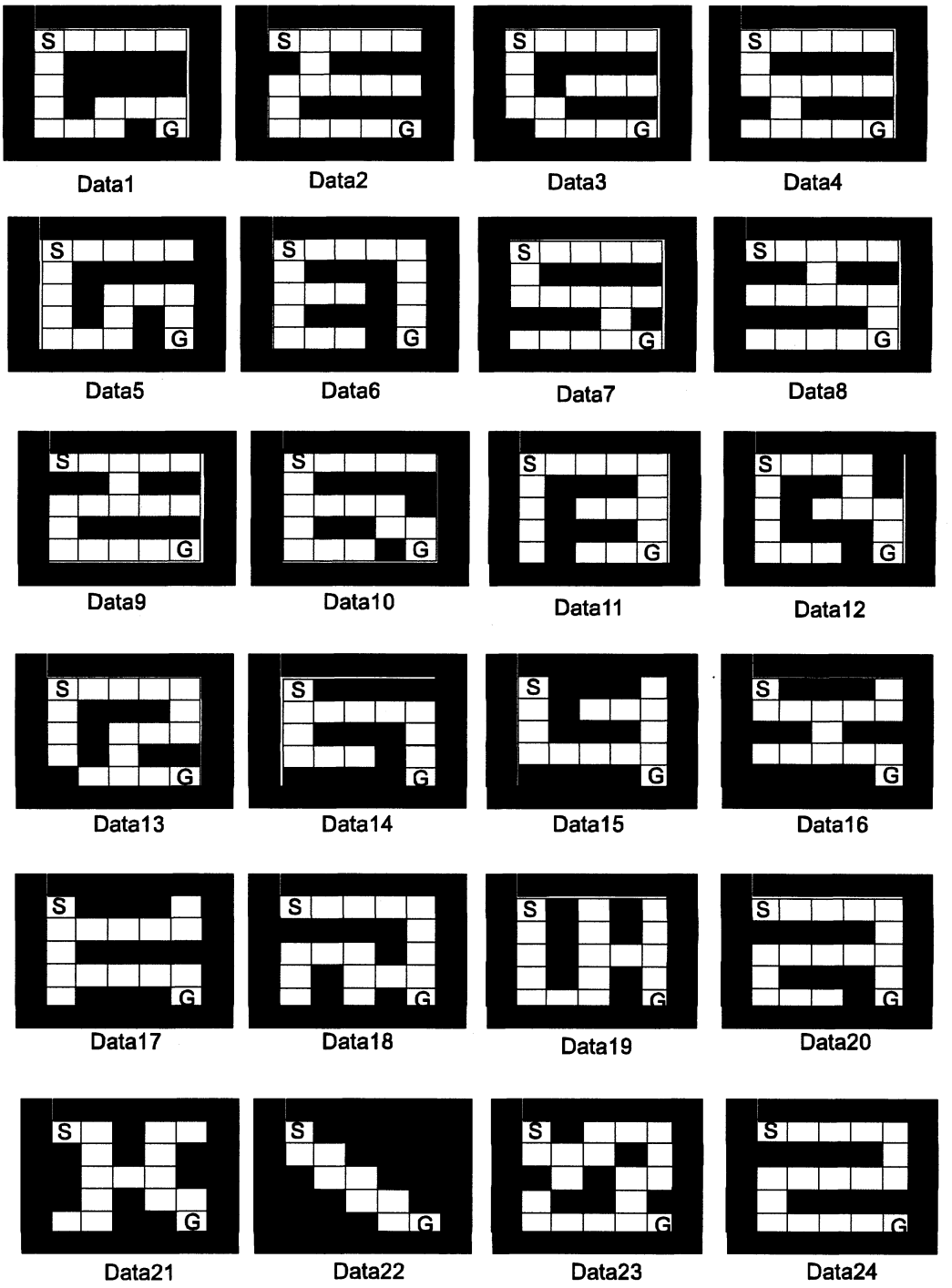

図 5.2: 過去に学習したタスク

5.2 選択された過去のタスク

前章で提案した転移させる知識の選択法において，どの知識が転移させる知識として選ばれるかを検証した．表5.1は提案法によって選ばれた過去のタスクの上位3位までを示している．

表 5.1: 選択されたタスク

タスク	順位	過去のタスク	選択回数 [trial]	学習達成度	平均の学習達成度
A	1	10	16	0.56	0.43
	2	5	3	0.76	
	3	6	3	0.06	
B	1	21	26	0.79	0.21
	2	-	-	-	
	3	-	-	-	
C	1	15	15	0.33	0.27
	2	19	3	0.33	
	3	13	1	0.27	

ここで，学習達成度は現在のタスクの通路マスの数，行うべき行動が現在の通路と過去の通路において同じマスの数，この二つの比によって算出したものである．この学習達成度により，過去の知識がどれだけ現在のタスクと一致しているかを表している．

表5.1より，提案法によって選ばれた過去のタスクは過去のタスクの中で，平均の学習達成度以上のものが主に選ばれていることが分かる．一部の例外として，タスクAでは学習達成度の低いものが選ばれてしまったが，選ばれた回数は少ない．以上により，禁止行動規則を用いて過去のタスクの知識を選ぶ方法は，学習達成度の高い過去のタスクを選ぶことができる．したがって，この手法により選ばれた過去の知識を転移させることによって現在のタスクの学習回数を削減できると考えられる．

5.3 提案手法による学習回数削減の効果

提案手法によって学習回数削減の効果が得られるかどうかを，5.1節で説明した各手法の学習回数を比較することで提案法の有効性を検証する．タスクA，B，Cにおいてそれぞれの手法を用いて，学習を行った結果をそれぞれ表5.2，表5.3，表5.4に示す．

それぞれの表は，5.1節の各手法における学習回数[episode]を表している．ただし，()内は学習失敗数[trial]を表している．また，太字はt検定(有意水準$p < 0.05$)により，Normalと比較し，有意に異なる値であることが確認された手法を表している．ただし，学習失敗数が2割(6[trial])以上となった手法については悪い結果とし，t検定を行っていない．また，NAは一度もタスクを達成することができなかったことを意味している．

表5.2: タスクAにおける各手法の学習回数

タスクA			p		
			転移あり		
			p_{pro}	p_{all}	転移なし
V	転移あり	V_{pro}	134.5(1)	153.1(28)	**102.0**(0)
		V_{all}	**212.5**(3)	**201.9**(3)	**288.3**(2)
	転移なし		108.6(9)	72.1(20)	140.4(0)

表5.3: タスクBにおける各手法の学習回数

タスクB			p		
			転移あり		
			p_{pro}	p_{all}	転移なし
V	転移あり	V_{pro}	**81.7**(0)	NA(30)	92.9(0)
		V_{all}	202.0(6)	**188.1**(0)	229.9(12)
	転移なし		157.5(1)	51(29)	108.1(0)

表 5.4: タスクCにおける各手法の学習回数

タスクC			p		
			転移あり		
			p_{pro}	p_{all}	転移なし
V	転移あり	V_{pro}	78.2(0)	135.2(16)	**142.2**(0)
		V_{all}	**156.4**(2)	**172.1**(2)	**176.3**(0)
	転移なし		**57.5**(1)	68.3(23)	95.4(0)

5.4 考察

表5.2，表5.3，表5.4より，提案手法はNormalより，現在のタスクやデータベースにある過去のタスクの知識によって，学習回数の削減が期待できることがいえる．なぜなら，提案手法はタスクBにおいては学習回数を有意に削減することができ，タスクA，タスクCにおいては有意に学習回数が増加しなかったためである．また，提案手法より，学習回数を削減できる手法が存在したが，同一の手法において，学習回数が増加してしまうタスクもあった．つまり，提案手法は，本論文で比較した手法の中で，最も安定して学習回数を削減できたといえる．

提案手法により学習回数を削減できた理由について考察する．禁止行動規則を用いたことにより選ばれた過去のタスクは5.2節でも示したように，現在のタスクと行動規則が多く一致する過去の知識が選ばれた．さらにいえば，スタートに近い部分が一致するものが多かった．これは，学習中に得られる禁止行動規則はスタートから徐々にゴールに近づくように禁止行動規則が得られることを意味しており，禁止行動規則を用いて転移する過去のタスクを選択する場合，スタートに近い部分が一致するものが多く選ばれ易い性質がある．このことから，行動優先度の転移については，過去のタスクと現在のタスクで禁止行動規則が一致する部分を利用したことで，スタートに近い部分について禁止行動をとることが少なくなったことが考えられる．また，表5.2，表5.3，表5.4において，p_{all}とV_{all}のどちらか一方でも利用した転移法は学習回数が増加している．このことから，それぞれの知識をそのまま利用することで，間違った知識まで利用してしまい，この知識の修正に多くの学習回数を必要となったと考えられる．提案法では，学習回数の増加がなかったことより，学習回数を増加させないように過去の知識を転移

できたと考えられる．以上より，提案法を用いて転移学習を行うことにより，現在のタスクの学習回数を削減できることを確認した．

第6章　まとめ

本研究では，アクター・クリティックにおいて転移学習によって学習回数を削減することを目的とした．転移学習の方法として，学習中に転移する過去の知識を決定し，学習回数削減の効果が高くなるように，行動優先度と状態価値のそれぞれの特徴をふまえて転移させる方法を検討した．

そのため，強化学習における報酬に着目し，負の報酬が与えられる行動（禁止行動）を収集し，禁止行動規則を定めた．これ用いて，学習中に転移させる過去の知識の決定が行える枠組みを提案した．また，アクター・クリティックにおける行動優先度，状態価値の性質と禁止行動規則を用いた転移させる知識の決定法の性質，2つの性質をふまえて，行動優先度は禁止行動が一致する状態のみを転移させ，状態価値は正の状態価値のみを転移させることを提案した．迷路問題を対象として，提案法の有効性を確認するための実験を行った．その結果，提案法を用いることで，学習回数を削減できることを確認した．

今後の課題としては，状態数を増やし，現実問題にも対応できるようにすること．あるいは，状態数の異なるエージェントへの転移学習を検討するなどが挙げられる．

謝辞

本論文は，著者が三重大学大学院工学研究科博士前期課程時に行った研究をまとめたものである．本論文を進めるにあたり，懇切丁寧な御指導と御督励を賜った三重大学の林照峯教授，鶴岡信治教授，北英彦准教授，高瀬治彦准教授，川中助教に感謝いたします．また，日頃熱心に討論していただいた計算機工学研究室，情報処理研究室の皆様方に厚く御礼申しあげます．

最後に，本論文をまとめるにあたり，助言，討論，その他お世話になったすべての方々に感謝いたします．

参考文献

[1] Richard S. Sutton, Andrew G. Barto 著, 三上貞芳, 皆川雅章 訳: 強化学習, 森北出版, 2000.

[2] Jeffery A. Clouse and Paul E. Utgoff: A Teaching Method for Reinforcement Learning, Proceedings of the Ninth International Workshop on Machine Learning, pp.92-110, 1992

[3] 荒井幸代: マルチエージェント強化学習：実用化に向けての課題・理論・諸技術との融合, 人工知能学会誌, Vol.16, No.4, pp.476-481, 2001.

[4] Witten, I. H.: An adaptive optimal controller for discretetime Markov environments. Information and Control, 34, pp.268-295, 1977.

[5] 亀井圭史, 石川眞澄: 移動ロボットの強化学習パラメータの環境依存性, 電子情報通信学会技術研究報告, NC, 102 巻 628 号, Vol.105, No.659, pp.61-66, 2006.

[6] 鮫島和行, 片桐憲一, 銅谷賢治, 川人光男: 複数の予測モデルを用いた強化学習による非線形制御, 電子情報通信学会論文誌, Vol.J84-D-II, No.9, pp.2092-2106, 2001.

[7] S. J. Pan and Q. Yang.: A survey on transfer learning. Technical Report HKUST-CS08-08, Dept. of Computer Science and Engineering, Hong Kong Univ. of Science and Technology, 2008.

[8] 涌田和芳: 不完全状態観測のセミマルコフ決定過程, 日本オペレーションズ・リサーチ学会論文誌, Vol.24, No.5, pp.95-109, 1981

[9] Lovejoy, W. S.: A Survey of Algorithmic Methods for Partially Observed Markov Decision Processes, Anuals of Operations Research 28, pp.47-65, 1991

[10] VIJAY R. KONDA AND JOHN N. TDITDIKLIS: ON ACTOR-CRITIC ALGORITHMES, Society for Industrial and Applied Mathematics J. CONTROL OPTIM. Vol. 42, No.4, pp.1143-1168, 2003

[11] 高玉圭樹: マルチエージェント学習 –相互作用のなぞに迫る–, コロナ社, 2003.

[12] R. Caruana. : Multitask learning, Machine learning, Vol.28, pp.41-7, 1997.

[13] Hoshino Y., Kamei K.: A Proposal of Reinforcement learningsystem to Use Knowledge Effectively, SICE Annual Conference 2003, Vol.2, pp.1582-1585, 2003.

[14] Fernando Fernandez, Manuela Veloso: Probabilistic Policy Reuse in a Reinforcement Learning Agent, Proceedings of the fifth International Joint Conference on Autonomous Agents and Multiagent Systems, pp.720-727, 2006.

[15] 松井藤五郎, 犬塚信博, 世木博久, 伊藤英則: 強化学習結果の再構築への概念学習の適用, 人工知能学会論文誌, Vol.17, No.2, pp.135-144, 2002.

[16] Leslie Pack Kaelbling, Michael L. Littman, Andrew W. Moore: Reinforcement Learning – A Survey, Journal of Artificial Intelligence Research, Vol.4, pp.237-285, 1996.

[17] 尾川 順子, 並木 明夫, 石川 正俊: 学習進度を反映した割引率の調整, 電子情報通信学会技術研究報告. NC, pp.73-78, 2003.

[18] Manu Sharma, Michael Holmes, Juan Santamaria, Arya Irani, Charles Isbell, Ashwin Ram: Transfer Learning in Real-Time Strategy Games Using Hybrid CBR/RL, Proceeding of International Joint Conference on Artificial Intelligence, pp.1041-1046, 2007.

[19] S. Thrun.: Is learning the n–th thing any easier than learning the first?, In Advances in Neural Information Processing Systems 8, pp.640-646, 1996.

発表論文

[1] 今井拓真, 高野敏明, 森田直樹, 高瀬治彦, 北英彦, 林照峯: 記述式小テストの解答の途中経過を講師に提供するシステム, 2008PC カンファレンス論文集, pp.228-231, 2008 (最優秀学生論文賞受賞)

[2] 高野敏明, 高瀬治彦, 北英彦, 林照峯: Actor-Critic における知識の再利用に関する一考察–再利用すべき知識の特徴抽出に関する一試み–, 第 26 回東海ファジィ研究会講演論文集, pp.18–1-18–4, 2009

[3] 高野敏明, 高瀬治彦, 川中普晴, 鶴岡信治: Actor-Critic における知識の再利用に関する一考察–特徴抽出と知識の再利用方法に関する一試み–, 第 27 回東海ファジィ研究会講演論文集, pp.9–1-9–4, 2009

[4] Toshiaki Takano, Haruhiko Takase, Hiroharu Kawanaka, Hidehiko Kita, Terumine Hayashi, Shinji Tsuruoka: Detection of the effective knowledge for knowledge reuse in Actor-Critic, Proceedings of the 19th Intelligent System Symposium and the 1st International Workshop on Aware Computing, pp.624-627, 2009

[5] Toshiaki Takano, Haruhiko Takase, Hiroharu Kawanaka, Hidehiko Kita, Terumine Hayashi, Shinji Tsuruoka: For Knowledge Reuse in Actor-Critic Detection of the Effective Knowledge, Proceedings of the 1st International Workshop on Regional Innovation Studies, pp.63-66, 2009