

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2007-66242

(P2007-66242A)

(43) 公開日 平成19年3月15日(2007.3.15)

(51) Int. Cl. F I テーマコード (参考)  
 G06N 3/00 (2006.01) G06N 3/00 550E 5H004  
 G05B 13/02 (2006.01) G05B 13/02 L

審査請求 未請求 請求項の数 3 O L (全 8 頁)

(21) 出願番号 特願2005-254763 (P2005-254763)  
 (22) 出願日 平成17年9月2日(2005.9.2)

特許法第30条第1項適用申請有り 2005年8月22日 社団法人情報処理学会発行の「FIT2005 第4回 情報科学技術フォーラム 一般講演論文集 第2分冊」に発表

(71) 出願人 504182255  
 国立大学法人横浜国立大学  
 神奈川県横浜市保土ヶ谷区常盤台79番1号  
 (74) 代理人 100095810  
 弁理士 鈴木 雄一  
 (72) 発明者 濱上 知樹  
 神奈川県横浜市保土ヶ谷区常盤台79番1号 国立大学法人横浜国立大学内  
 (72) 発明者 ▲渋谷 長史  
 神奈川県横浜市保土ヶ谷区常盤台79番1号 国立大学法人横浜国立大学内  
 Fターム(参考) 5H004 GA02 GB16 HA07 HB07 HB08  
 JA03 JB01 KD61

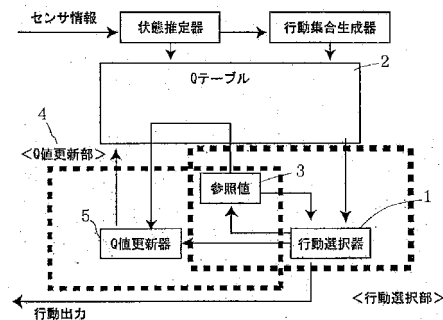
(54) 【発明の名称】 強化学習の価値関数表現方法およびこれを用いた装置

(57) 【要約】

【課題】強化学習法は自律移動ロボット等に应用される知的動作の一つであり、未知の環境での動作が可能であるなど優れた面を持つシステムであるが、不完全知覚問題という基礎的な課題を持っている。いろいろな解決策が提案されているが、決め手はなく、システムも複雑なものになり兼ねない。簡単で効果的な解決方法が望まれている。

【解決手段】状態行動価値を複素数で定義する複素価値関数を導入する。時系列情報は複素数値の位相部分に取り込むことにする。これにより、複雑なアルゴリズムを用いることなく時系列情報が価値関数に取り込まれ、簡便な実装により不完全知覚問題が有効に解決される。

【選択図】 図4



## 【特許請求の範囲】

## 【請求項 1】

価値関数に複素数値関数を使用し、時系列情報を複素数の位相部分もしくは位相部分および振幅部分に算入することにより、不完全知覚問題に対応することを特徴とする強化学習の価値関数表現方法。

## 【請求項 2】

前記強化学習法として Q ラーニング法を用いたことを特徴とする特許請求範囲 1 に記載の強化学習の価値関数表現方法。

## 【請求項 3】

価値関数に複素数値関数を使用し、時系列情報を複素数の位相部分もしくは位相部分および振幅部分に算入することにより、不完全知覚問題に対応することを特徴とする強化学習の価値関数表現方法を用いた装置。

10

## 【発明の詳細な説明】

## 【技術分野】

## 【0001】

本発明は、自律移動ロボット等に応用する強化学習方法およびこれを用いた装置に関する。

## 【背景技術】

## 【0002】

強化学習は自律移動ロボットのような行動主体が、自ら環境を観測し行動した結果から次の適切な方策を獲得する知的動作である。とくに環境同定型手法は教師信号を使わない学習手段であるため、未知の環境における行動を決めるのに向いたシステムであると言われている。代表的な強化学習方法として、Q ラーニングをはじめとする状態行動対の価値関数を求める環境同定型手法や、メモリに記憶したエピソードを利用する経験強化型手法が知られている。

20

## 【0003】

強化学習法の一般理論については [1] S. Russell and P. Norvig: Artificial Intelligence A Modern Approach, Prentice Hall, 1995 (邦訳「エージェントアプローチ 人工知能」共立出版 1997) または [2] R. S. Sutton and A. G. Barto: Reinforcement Learning An Introduction, The MIT Press 1988. (邦訳「強化学習」森北出版、2000) が詳しい。

30

強化学習法についての改良、応用は数多くあり、例えば、基本アルゴリズムに関して言えば、学習における連続状態空間の扱いや、学習速度向上を目指した研究開発が行われている。例えば、[3] エージェント学習装置 (科学技術振興事業団、特許文献 1) がある。

## 【0004】

## 【特許文献 1】特開 2000 - 35956

## 【発明の開示】

## 【発明が解決しようとする課題】

## 【0005】

強化学習法における基礎的な問題に「不完全知覚問題」がある。環境同定型強化学習では状態と行動の対に対して価値関数の値を決める。この値が大きいほどその状態ととるべき行動としてふさわしいとするのである。アルゴリズムが比較的簡単で実装が容易である反面、現実の環境と環境検出能力では、ある状態に該当する空間が無数に存在し、その度に異なる行動の選択が求められるという問題が生ずる。これが「不完全知覚問題」である。

40

## 【0006】

不完全知覚問題の生ずる環境は非マルコフ過程からなる環境であり、Q ラーニングをはじめとする従来からの強化学習法では原理的に対応できないことが知られている。不完全知覚問題については設計者のヒューリスティクスに基づく仮定や、新しいパラメータを導入することが考えられたが、効果が保障されたわけではない。例えば、前出の文献 [3] では、環境の変化を予測し、変化に応じて複数の学習モジュールを自動的に切り替える方

50

法を提供しているが、各学習モジュールの分担範囲を決めるパラメータはタスクに依存するという問題がある。

【0007】

文献[4]特開2005-78519内部変数推定装置、内部変数推定方法及び内部変数推定プログラム((株)国際電気通信基礎技術研究所)では、内部変数を推測する機構を有し、直接観測できない状態をメタパラメータとして表す方法を提案しているが、内部状態の数や内部変数の次元の設定は設計者のヒューリスティクスに依存する。また、文献[5]特開平9-81205学習システム(富士通(株))によれば、いくつかの時系列情報(コンテキスト)をメモリに蓄え、状態の履歴からとるべき行動を決定するエピソード記憶方式の経験強化型強化学習方法が提案されている。この方法はコンテキストを直接保持するため、非マルコフ過程の環境に対応できるが、どれだけの長さのコンテキストを持てばよいのか、学習時の探索範囲の設定など、設計者のヒューリスティクスに大きく依存せざるを得ない。信頼性の点で問題の多い、ヒューリスティクスに依存せず、かつ、メモリその他の資源を多大に使わない解決策が望まれる所以である。

10

【課題を解決するための手段】

【0008】

本発明における問題解決の要点はエピソード記憶のようなコンテキストを価値関数に簡便な方法で取り込むことにある。このために状態行動価値を複素数で定義する複素価値関数を導入する。時系列情報は複素数値の位相部分に主として取り込まれる。これにより、複雑なアルゴリズムを用いることなく時系列情報が価値関数に取り込まれ、容易な実装でありながら、不完全知覚問題が解決できることとなる。

20

【0009】

すなわち、請求項1記載発明は、価値関数に複素数値関数を使用し、時系列情報を複素数の位相部分もしくは位相部分および振幅部分に算入することにより、不完全知覚問題に対応することを特徴とする強化学習の価値関数表現方法である。また請求項2記載発明は前記強化学習法としてQラーニング法を用いた強化学習の価値関数表現方法である。

【0010】

請求項3記載発明は、価値関数に複素数値関数を使用し、時系列情報を複素数の位相部分もしくは位相部分および振幅部分に算入することにより、不完全知覚問題に対応することを特徴とする強化学習の価値関数表現方法を用いた装置である。

30

【発明の効果】

【0011】

本発明は不完全知覚問題を複雑なアルゴリズムを用いることなく簡便な実装で解決するものであるから、不完全知覚問題のもたらす本質的な欠陥が解消し、自律移動学習が可能なロボットが容易に作れるようになる。不完全知覚問題の及ぶ範囲は広大であり、本発明が解決する問題の範囲も自ずから広いものとなり、技術的、経済的効果は多大である。

【発明を実施するための最良の形態】

【0012】

具体例として、QラーニングにおけるQ値を複素数として扱う方法を説明する。Q値が複素数であることを明示的に複素Q値と表す。複素Q値の更新式において遷移先の状態に関連する複素Q値をとる際に、位相回転を加えることで時系列の情報(コンテキスト)を含ませるのが本発明の要点である。すなわち直前の行動の複素Q値との位相差を考慮して次のステップで選択されるであろう複素Q値を予測する。図1は予測された複素Q値(複素ベクトルR)と選択可能な行動に対応する複素Q値(複素 $Q_1$ 、複素 $Q_2$ )の関係を示す。複素ベクトルRの位相項が変化すると各複素Q値との内積も変化する。つまり、各複素Q値から複素ベクトルRと原点を結ぶ直線に直角に下ろした足と原点との長さが変化する。Q値を実数として扱う場合は、単純に大きさの比較を行うことしか出来ないが、複素Q値を用いると位相差を含んだ比較が可能となる。これによって複雑なアルゴリズムを使わずに時系列を取り入れた行動選択が可能になる。

40

【0013】

50

[更新アルゴリズムの定式化]

状態  $s_i$  から行動  $a_i$  をとって状態  $s_{i+1}$  へと遷移し報酬  $r$  を受け取ったときの、複素  $Q$  値の更新則を数 1 のように定義する。

【0014】

【数 1】

$$\dot{Q}(s_{i-k}, a_{i-k}) \leftarrow (1 - \alpha)\dot{Q}(s_{i-k}, a_{i-k}) + \alpha(r + \gamma\dot{Q}_{max}^{s_i \rightarrow s_{i+1}})\dot{u}(k)$$

【0015】

10

ここで、 $k$  ステップ前の状態、行動をそれぞれ  $s_{i-k}$ ,  $a_{i-k}$  とする。 $u(k)$  は複素関数であり、形式上の適格度トレースであり、数 2 のように定義する。数 2 中では関数  $u$  に複素数を示すドットを付けた。数 2、5 中では関数  $\dot{u}$  に複素数を示すドットを付けた。

【0016】

【数 2】

$$\dot{u}(k) = \dot{\beta}^k$$

20

数 1 の適用は、予め定めた整数  $N_e$  を用いて、 $0 \leq k \leq N_e$  の範囲で行う。

ただし、 $\dot{\beta}$  は絶対値が 1 以下の複素数である。

数 1 における複素  $Q$  値は数 3 のように定義する。

【0017】

【数 3】

$$\dot{Q}_{max}^{s_i \rightarrow s_{i+1}} = \dot{Q}(s_{i+1}, \dot{a})$$

30

ただし、 $\dot{a}$  は数 4 のように定義する。

【0018】

【数 4】

$$\dot{a} = \arg \max Re[\dot{Q}(s_{i+1}, \dot{a})\overline{R_i}]$$

40

ここで、予想される複素  $Q$  値 (複素ベクトル  $R_i$ ) は、数 5 のように定義する。

【0019】

【数 5】

$$R_i = \dot{Q}(s_i, a_i) / \dot{\beta}$$

50

## 【0020】

## [ 行動選択アルゴリズムの定式化 ]

ここでは、Max-Boltzmann 選択を用いる。すなわち、状態  $s_i$  に居るエージェントは、確率  $1 - P_{max}$  で Boltzmann 選択を行い、確率  $P_{max}$  で Greedy 方策を行うことにする。

状態  $s_i$ 、行動  $a_i$  に対応する複素  $Q$  値を複素  $Q(s_i, a_i)$  とする。また、状態  $s_i$  における行動  $a$  の Boltzmann の選択確率を  $Prob(s_i, a)$  とする。状態  $s_i$  における行動集合を  $A(s_i)$ 、直前の状態と行動に対応する複素  $Q$  値を複素  $Q(s_{i-1}, a_{i-1})$ 、Boltzmann 選択の温度パラメータを  $T$  とするとき、 $Prob(s_i, a)$  を数 6 のように定める。

10

## 【0021】

## 【数 6】

$$Prob(s_i, a) = \frac{\exp(\operatorname{Re}[\dot{Q}(s_i, a)\overline{R_{i-1}}]/T)}{\sum_{a \in A(s_i)} \exp(\operatorname{Re}[\dot{Q}(s_i, a)\overline{R_{i-1}}]/T)}$$

ただし、 $\operatorname{Re}$  [ 複素関数 ] は複素数の実部を表す。

greedy 方策は  $\arg \max_a Prob(s_i, a)$  を選択することにする。

20

## 【0022】

## [ 計算機実験 ]

図 2 のような簡単なグリッドワールドにおける迷路問題を対象として計算機実験を行い、提案手法の有効性を確認する。

## 【0023】

## [ 状態空間と行動集合 ]

エージェントが観測可能な情報は、東西南北周囲 4 マスの壁の有無のみとし、この情報を直接状態として割り当てることにする。すなわち観測可能な状態数は  $2^4 = 16$  となる。これらの環境において不完全知覚の影響のある状態が存在する。例えば、アスタリスク \* においてはそれぞれにおいて選択すべき行動が異なり、\*\* においては同じ行動をとらなければならない。エージェントが任意の状態において選択することができる行動は、壁のない方向に進むのみとする。すなわち、行動集合  $A = \{ \text{東、西、南、北} \}$  の空集合でない部分集合とする。

30

## 【0024】

## [ パラメータ設定 ]

エージェントは、ゴールにたどり着くと環境から報酬  $r = 100$  を受け取り、初期状態であるスタートに再配置されるものとした。エージェントの行動 1 ステップごとに負の報酬を与えることや、ゴールにたどり着くのかにかかったステップ数に応じて報酬を変えることなど、早くゴールにたどり着く学習を助長するような報酬の与え方はしない。

試行数 100 を 3 つのフェーズに分け、それぞれについてパラメータの設定を行った。ステップごとに変化するパラメータについては表 1 のように設定し、それ以外のパラメータについては各フェーズにおいて共通とし、 $\beta = 0.9 \exp(j/6)$ 、 $\gamma = 0.999$ 、 $T = 3000$ 、 $N_e = 1$  とした。ただし、 $j^2 = -1$  である。

40

## 【0025】

【表 1】

	試行	$\alpha$	$1-P_{\max}$
フェーズ 1	1to20	0.05	0.1
フェーズ 2	21to80	$(100-\text{try})/400$	$(100-\text{try})/1600$
フェーズ 3	81to100	0	0

## 【0026】

## [実験結果]

計算機実験の結果を図3に示す。この結果は100試行を1学習として100学習行い、収束したものに關しての平均である。

maze 1、maze 2においては100%が収束し、maze 3においては95%が収束した。本計算機実験ではmaze 1、maze 2において100%が最短経路を実現するような方策を獲得した。

## 【0027】

maze 1では最短経路を実現する方策が獲得でき、一連の行動について観察すると、ある複素Q値次の行動に対応する複素Q値と の偏角だけずれる学習がなされている。maze 2でも最短経路を実現する方策が獲得できた。maze 1のような単純な位相関係ではなかったが、位相を自律的に調整することで、不完全知覚問題を解決していることが観察された。maze 3では最短経路を実現する方策の学習は見られなかったが、環境中を一部往復することで自律的に環境を多重化して不完全知覚問題を解決していることが観察された。

いずれの場合も、問題を自律的に解決する行動が獲得できており、本発明による不完全知覚問題の解決の効果が示されている。

## 【0028】

図4は本発明に關わる装置の具体例で、複素Qラーニング法を実装した装置のブロック図である。行動選択器1においては前回の参照値を基準としてQテーブル2から渡されるQ値集合の中から一つの値を選択する。Q値変更部4では新たな参照値を基準として遷移後のQ値集合の中から一つの値を選択して更新の目標値とし、Q値変更器5で変更を実行する。

## 【0029】

なお、本実施例では価値関数の複素数化と位相の取り込み方を、[数2]のように計算したが、複素数化と位相の取り込み方はこれに限られるものではない。例えば、図1において複素 $Q_1$ 、複素 $Q_2$ の位相を時系列情報に基づいて変化させたり、位相だけでなく振幅を変化させてもよく、実際の計算法は環境によって適宜選択されるべきである。本発明の主旨は複素価値関数を用い、位相項に時系列情報を取り入れることにあり、いかなる取り込み方にも及ぶものである。

## 【0030】

また、本実施例では複素価値関数を用いる手法をQラーニング法に応用した例を示したが、本発明の本質は複素価値関数を用いることにあるので、例えば、TD法、SARSA法、Actor-Critic法、R学習法、Profit-Sharing法などの価値関数を用いる方法であればいずれも有効に機能する。

## 【0031】

本実施例では行動選択アルゴリズムとしてMax-Boltzmann選択を採用したが、複素価値関数の出力である複素数値から実数値の選択確率を計算できるアルゴリズムであればどのようなものでも良く、理工学で一般的に使われているもので計算する方法は何れも本発明の範囲内に入るものである。

## 【0032】

産業上の利用可能性：

10

20

30

40

50

不完全知覚問題は強化学習の基本的問題であり、本発明によりこれが解決されれば環境同定型学習の多くの問題点が自ずから解消し、廉価なセンサを有するロボットで自立移動学習が可能になる。本発明の実装は廉価容易であり、経済的な効果は大きい。さらに不完全知覚問題はマルチエージェント系など多数の学習主体の同時学習でも現れる問題であり、本発明はマルチロボットや多点探査アルゴリズムなど、集団としての効率的な学習を要求される用途にも有効に使用できる。

【図面の簡単な説明】

【0033】

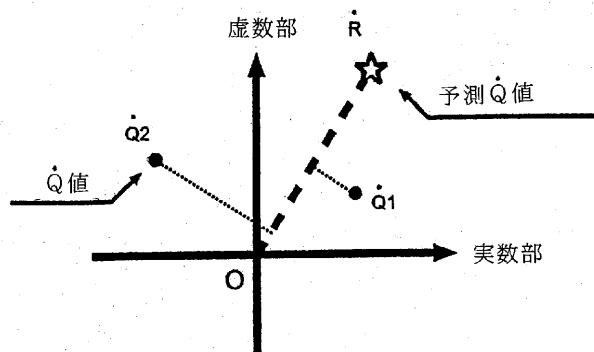
【図1】ある状態における複素Q値を複素平面上に示した図である。

【図2】(a)～(c)は計算機実験の実験環境を示す図で、簡単な迷路問題を行うグリッドワールドを示す図である。 10

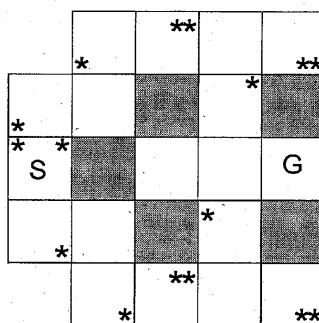
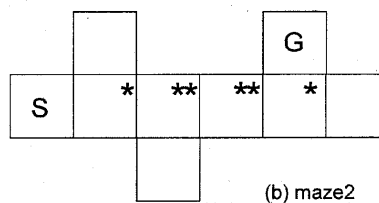
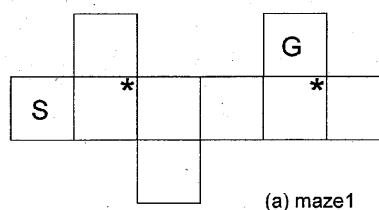
【図3】計算機実験の結果を示す図である。

【図4】本発明にかかわるQラーニング法を実装した装置のブロック図である。

【図1】

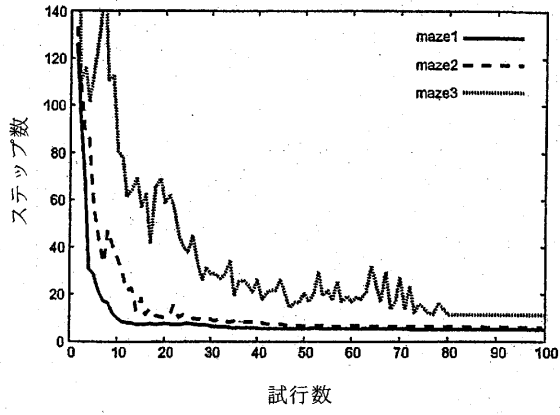


【図2】



(c) maze3

【 図 3 】



【 図 4 】

