(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号
特許第4929449号
(P4929449)
(45)発行日　平成24年5月9日(2012.5.9)　(24)登録日　平成24年2月24日(2012.2.24)

(51)Int.Cl.　　F I
G06N 3/00 (2006.01)　　G06N 3/00 550E
G05B 13/02 (2006.01)　　G05B 13/02 L

請求項の数 4　(全 9 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)
(65)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)
審査請求日　平成20年9月1日(2008.9.1)

特許法第３０条第１項適用　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

(73)特許権者　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　100094053
弁理士　佐藤　隆久
(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

審査官　新井　寛

最終頁に続く

(54)【発明の名称】強化学習装置および強化学習方法

(57)【特許請求の範囲】
【請求項１】
状態と行動の対に対して価値関数値を決定し保持する価値関数値保持部と、
前記価値関数保持部から渡される価値関数値集合の中から１つの値を選択し、選択した値を基に行動を選択する行動選択部と、
状態が遷移した時に価値関数値を更新する価値関数値更新部と、
を有し、
前記価値関数値は複素数であり、
前記価値関数値更新部は、前記価値関数値の更新式において、直前の行動の複素価値関数値との位相差を考慮して複素価値関数値を更新する
強化学習装置。
【請求項２】
前記行動選択部は、前記選択した価値関数値と、前記直前の行動の複素価値関数値を基に算出された複素ベクトルの共役複素数との積の実部を使用して行動を選択する
請求項１に記載の強化学習装置。
【請求項３】
前記強化学習装置における強化学習方法として、Ｑラーニング法を用い、
前記価値関数値更新部は、前記価値関数値の更新式において、適格度トレースアルゴリズムを使用する
請求項１又は２に記載の強化学習装置。

【請求項４】
価値関数値保持部と、行動選択部と、価値関数値更新部と、を有する強化学習装置の強化学習方法であって、
前記価値関数値保持部が、状態と行動の対に対して価値関数値を決定し保持する第１のステップと、
前記行動選択部が、前記価値関数保持部から渡される価値関数値集合の中から１つの値を選択し、選択した値を基に行動を選択する第２のステップと、
前記価値関数値更新部が、状態が遷移した時に価値関数値を更新する第３のステップと、
を有し、
前記第１のステップにおいて、前記価値関数値保持部が決定する価値関数値は複素数であって、
前記第３のステップにおいて、前記価値関数値更新部は、前記価値関数値の更新式において、直前の行動の複素価値関数値との位相差を考慮して複素価値関数値を更新する
強化学習装置の強化学習方法。
【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。
【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。
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で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。
【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イッ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イックスに依存する。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、学習時の探査範囲の設定など、設計者のヒューリステッ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、ヒューリステックスに依存せず、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。

【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。

【０００９】

すなわち、本発明の強化学習装置は、状態と行動の対に対して価値関数値を決定し保持する価値関数値保持部と、前記価値関数保持部から渡される価値関数値集合の中から１つの値を選択し、選択した値を基に行動を選択する行動選択部と、状態が遷移した時に価値関数値を更新する価値関数値更新部と、を有し、前記価値関数値は複素数であり、前記価値関数値更新部は、前記価値関数値の更新式において、直前の行動の複素価値関数値との位相差を考慮して複素価値関数値を更新する。

【００１０】

本発明の教科学習方法は、価値関数値保持部と、行動選択部と、価値関数値更新部と、を有する強化学習装置の強化学習方法であって、前記価値関数値保持部が、状態と行動の対に対して価値関数値を決定し保持する第１のステップと、前記行動選択部が、前記価値関数保持部から渡される価値関数値集合の中から１つの値を選択し、選択した値を基に行動を選択する第２のステップと、前記価値関数値更新部が、状態が遷移した時に価値関数値を更新する第３のステップと、を有し、前記第１のステップにおいて、前記価値関数値保持部が決定する価値関数値は複素数であって、前記第３のステップにおいて、前記価値関数値更新部は、前記価値関数値の更新式において、直前の行動の複素価値関数値との位相差を考慮して複素価値関数値を更新する。

【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ないで時系列を取り入れた行動選択が可能になる。

【００１３】

［更新アルゴニズムの定式化］

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】

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

【００１５】

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βに複素数を示すドットを付けた。

【００１６】

【数２】

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

数１の適用は、予め定めた整数$Ｎ_{e}$を用いて、$０ \leqq ｋ \leqq Ｎ_{e}$の範囲で行う。

ただし、β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】

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

ただし、ａ’は数４のように定義する。

【００１８】

【数４】

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

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】

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

【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ーをＴとするとき、Ｐｒｏｂ（ｓ$_{i}$，ａ）を数６のように定める。

【００２１】

【数６】

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

ただし、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。

【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＝｛東、西、南、北｝の空集合でない部分集合とする。

【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、β＝０．９ｅｘｐ（ｊπ／６）、γ＝０．９９９、Ｔ＝３０００、N_e＝１とした。ただし、$j^2=-1$である。

【００２５】

【表１】

	試行	α	$1-P_{max}$
フェーズ１	1to20	0.05	0.1
フェーズ２	21to80	(100-try)/400	(100-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、収束したものに関しての平均である。

ｍａｚｅ１、ｍａｚｅ２においては１００％が収束し、ｍａｚｅ３においては９５％が収束した。本計算機実験ではｍａｚｅ１、ｍａｚｅ２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】

ｍａｚｅ１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βの偏角だけずれる学習がなされている。ｍａｚｅ２でも最短経路を実現する方策が獲得できた。ｍａｚｅ１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。ｍａｚｅ３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は本発明に関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（本発明の価値関数値保持部に対応）から渡されるＱ値集合の中から一つの値を選択する。Ｑ値更新部４（本発明の価値関数値更新部に対応）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値更新器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：

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】

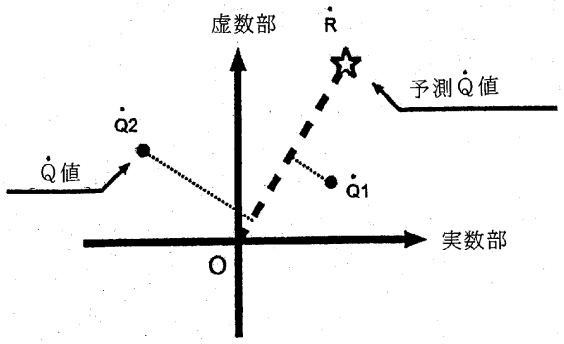
【図１】ある状態における複素Ｑ値を複素平面上に示した図である。

【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トワールドを示す図である。

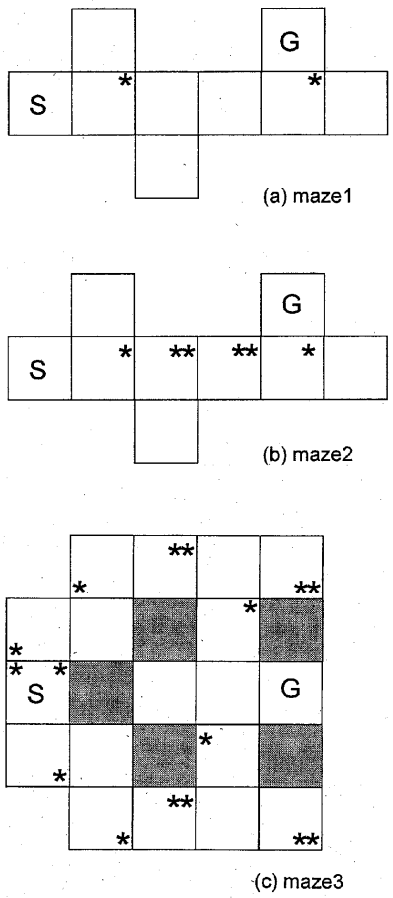
【図３】計算機実験の結果を示す図である。

【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。

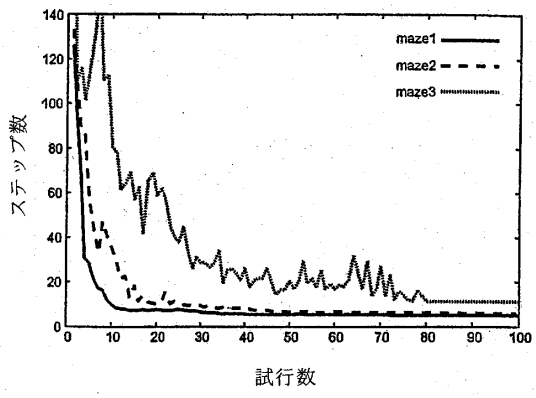
【図１】



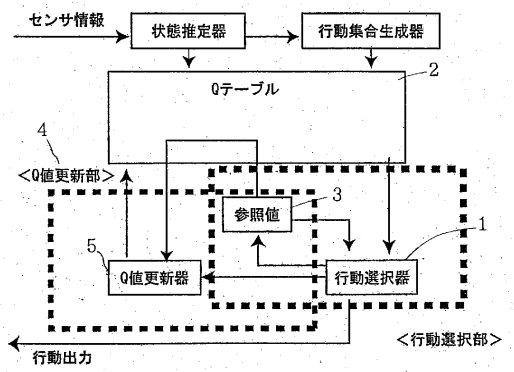
【図２】



【図３】


maze1
maze2
maze3
ステップ数
試行数
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60
70
80
90
100

【図４】


センサ情報
状態推定器
行動集合生成器
Qテーブル
2
4
<Q値更新部>
3
参照値
1
5
Q値更新器
行動選択器
<行動選択部>
行動出力

フロントページの続き

(56)参考文献 山本 真也，外３名，“強化学習における環境変化認識法”，電子情報通信学会技術研究報告，社団法人電子情報通信学会，２０００年 1月１３日，第９９巻，第５３４号，ｐ．３１－３６
植村 渉，外２名，“ＰＯＭＤＰｓ環境下での経験強化型強化学習法”，電子情報通信学会技術研究報告，社団法人電子情報通信学会，２００４年 7月２９日，第１０４巻，第２３３号，ｐ．１－５

(58)調査した分野(Int.Cl.，ＤＢ名)
Ｇ０６Ｎ 3／００
Ｇ０５Ｂ １３／０２