様式 C-19

科学研究費補助金研究成果報告書

平成 21 年 5 月 22 日現在

研究種目：若手研究（B）
研究期間：2006～2008
課題番号：18700145
研究課題名（和文）　エージェントの学習における適切な効用の導出方法

研究課題名（英文）　Deriving Appropriate Utility in Agent Learning

研究代表者
森山　甲一　(MORIYAMA KOICHI)
大阪大学・産業科学研究所・助教
研究者番号：10361776

研究成果の概要：
本研究課題は，人工知能研究の一分野である強化学習において，従来は同一視されてきた個体（エージェント）外からの報酬と，エージェント自身の主観的効用を分けて考えることで，最も単純なマルチエージェント環境である 2 人 2 行動同時手番対称ゲームで適切に行動する強化学習エージェントの構築を目的とした．研究の結果，囚人のジレンマゲームにおいて協調行動を続けやすくする効用の条件を導き，さらに別の種類のゲームで報酬を追求することを妨げない学習手法を開発した．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
2006 年度	800,000	0	800,000
2007 年度	800,000	0	800,000
2008 年度	700,000	210,000	910,000
年度			
年度			
総　計	2,300,000	210,000	2,510,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：人工知能，強化学習，エージェント，マルチエージェントシステム

１．研究開始当初の背景

我々人間は，他者の内面に直接アクセス出来ないという意味で疎結合な社会というネットワークを構成している．それは、あまりにも多くの個人が同時に存在し活動するため、個々の活動が完全には把握できないものであるが，人間はそのような社会に適応することが出来る．社会において適応する際に重要であると思われる点は，人間は必ずしも自己の利益のみを指向するのではなく，相互に譲り合うことによって望ましい状態を築くことであり，心理学実験などでも観察されている．

一方，人工知能研究においては，コンピュータ上に構築されたエージェントと呼ばれる仮想的な人工の個体において知能を実現しようとする研究が行われてきた．その中でも，エージェントが意思の決定と外部刺激（報酬と呼ぶ）の受領を繰り返すことにより，大きな報酬が得られる意思決定法を獲得する強化学習がエージェントの学習手法として広く研究され，実際に人間の脳における学習のモデルとしての議論も行われていた．

ところが，人間社会のモデルとして，エージェントを複数導入して仮想的な社会を構成した場合，報酬を最大化することを目的としたエージェントおよび強化学習では，各自が自己の報酬を最大化することを試み，エージェントが互いに足を引っ張る結果となることが観察された．

さらに一方では，人間において観察される譲り合いなどの協調行動がどのようなメカニズムで発生するのかについても明確な答えは得られていないため，それに基づいてエージェントを構築することも困難であった．

２．研究の目的

人間の適応性は，個人の観点では学習により獲得されるものである．学習とは，各個人が意思決定を繰り返すことで，各々の「嬉しさ」すなわち主観的効用を大きくする意思決定法を獲得することと言える．このように捉えると，報酬の最大化を行う強化学習手法を用いたエージェントでも同様のことを実現することが可能であると思われる．しかし，ここで問題となるのが，人間の主観的効用が単純には定義できない点である．人間の場合，効用には金銭などの外からの客観的報酬に加えて，個人の性格や他者との関係なども係わるが，強化学習の枠組みでは効用と報酬を同一視している．従って，社会に適応可能な強化学習エージェントを構築するには，効用と報酬を分けて考える必要があると思われる．

そこで，研究代表者は，これまで強化学習において顧みられなかった効用そのものに関する研究を考えついた．しかも，経済学などの研究で行われているような，政府などの権力者が税金や補助金などでエージェントの報酬を変えることによって行動を制御するのではなく，人間における個人の性格などの観点から，各エージェントが独自に効用を導出して強化学習を行なうこととし，その効用の導出方法を求めることを目的とした．

適応すべき社会というものはそれこそいろいろなものが考えられるが，本研究では，最も単純なマルチエージェント環境である非協力２人２行動同時手番対称ゲームのみを扱うこととした．ゲームに関する知識を全く持たない，同一の効用導出関数を持つ強化学習エージェント２台がそのゲームを繰り返し行なうものとした．特に，囚人のジレンマゲームと呼ばれるゲームでエージェントの協調を導く効用生成法を求めることを第一の目的とし，さらに他のゲームにおいても適切な行動を選択できるようにすることを第二の目的とした．

３．研究の方法

本研究は，数多くのシミュレーション実験を基盤とし，それから得られた興味深いデータから原因を調査し，その結果をもたらすための条件を理論的に検討するという流れで進められた．

(1) 非協力２人２行動同時手番ゲーム環境および強化学習の一種である Q 学習を用いたエージェントを計算機上に構築し，まずは囚人のジレンマゲームで，ゲームの利得表と学習に用いる効用のそれぞれについて少しずつ設定を変更したシミュレーション実験を数多く実行することにより，適切な効用導出関数についての情報を収集した．

(2) シミュレーション実験により得られた興味ある結果から，そのような結果をもたらすための条件を理論的に明らかにした．さらに，この条件を元に新たな効用導出関数を考案し、再びシミュレーションなどで検証した．ところが，この条件はエージェントが報酬を追求することを弱めるため，囚人のジレンマゲーム以外のゲームでは望ましい結果が得られないことが判明した．

(3) 上述の条件がゲームの利得と学習率の関数であったことから，効用にこだわらずに学習率を調整することによって，囚人のジレンマゲーム以外のゲームでも望ましい結果を得ることを目的とした研究を進めた．

４．研究成果

本研究の成果は大きく分けて以下の通りである．

(1) 囚人のジレンマゲームにおいて協調行動を導く効用関数の導出と，それを用いた強化学習手法（効用利用 Q 学習）の提案

(2) 囚人のジレンマゲームにおいて協調行動を導きつつ，それ以外のゲームにおいても望ましい結果を得る学習率の調整方法の導出と，それを用いた強化学習手法（学習率調整 Q 学習）の提案

Q 学習では，エージェントは自分の持つ行動価値関数（Q 関数）に基づいて行動を選択し，その行動の結果として得られた報酬を用いて Q 関数を更新することにより，選択した行動の価値（Q 値）を学習する．行動 a_t で報

酬 r_{t+1} を獲得した時の 1 状態 Q 学習の更新式は以下の通りである．$\alpha \in [0,1]$は学習率である．

$$Q_{t+1}(a) = \begin{cases} (1-\alpha)Q_t(a_t) + \alpha\delta_t & if\ a = a_t, \\ Q_t(a) & otherwise. \end{cases}$$

$$\delta_t = r_{t+1} + \gamma \max_a Q_t(a) - Q_t(a_t).$$

各エージェントが行動 C と D から選択を行うとき，非協力 2 人 2 行動同時手番対称ゲームは以下のように表される．

	C	D
C	r_{cc}, r_{cc}	r_{cd}, r_{dc}
D	r_{dc}, r_{cd}	r_{dd}, r_{dd}

但し，各エージェントは行または列から行動を選択するものとし，出現した行動の組み合わせから，該当するセルに含まれる利得（左側が行選択エージェント，右側が列選択エージェント）を獲得するものとする．

囚人のジレンマゲームでは，利得の順序が $r_{dc} > r_{cc} > r_{dd} > r_{cd}$ となる．この場合，相手の行動に係わらず D を選択した方が自分の利得が大きくなるが，両者が D を選択した場合よりも C を選択した場合の利得のほうが大きい．

効用利用 Q 学習は，両者が偶然 C を選択した場合に，C の Q 値を大きくすることで，それ以降も C を続けさせることが出来るものである．具体的には，両者が C を選択した場合に，利得 r_{cc} に以下の条件を満たす r を加えたものを効用として Q 学習を行う．

$$r \geq \frac{r_{dd} - (\alpha r_{cc} + (1-\alpha) r_{cd})}{\alpha}.$$

一方，学習率調整 Q 学習は，以下の式で表される学習率を利用して Q 学習を行う．但し，(X,Y)は自分の行動が X，相手の行動が Y であることを表す．α_0 は通常の学習率を表し，α_1, α_2, α_3 はそれぞれ下の条件を満たす．

$$\alpha = \begin{cases} \alpha_1 & if\ (C,C)\ \&\ Q_t(C) < Q_t(D), \\ \alpha_2 & if\ (C,D)\ \&\ Q_t(C) > Q_t(D), \\ \alpha_3 & if\ (D,C)\ \&\ Q_t(C) > Q_t(D), \\ \alpha_0 & otherwise. \end{cases}$$

$$\alpha_1 > \frac{Q_t(D) - Q_t(C)}{r_{t+1} + \gamma Q_t(D) - Q_t(C)},$$

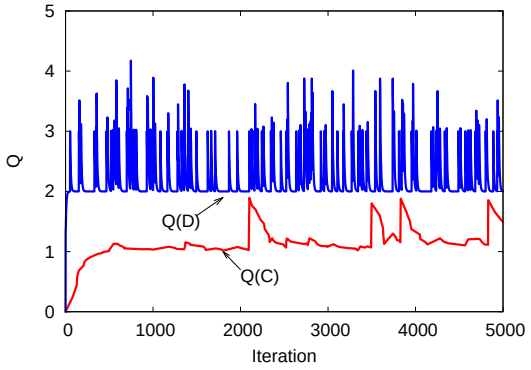
$$\alpha_2 < \frac{Q_t(C) - Q_t(D)}{(1-\gamma)Q_t(C) - r_{t+1}},$$

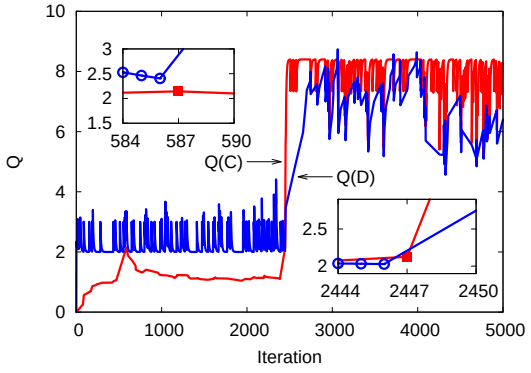
$$\alpha_3 < \frac{Q_t(C) - Q_t(D)}{r_{t+1} - Q_t(D) + \gamma Q_t(C)}.$$

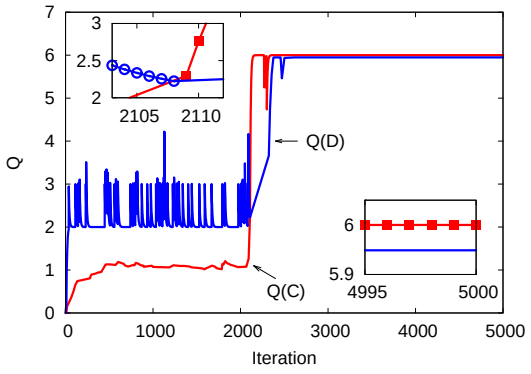
ここで，行き詰まりゲームを導入する．行き詰まりゲームは $r_{dc} > r_{dd} > r_{cc} > r_{cd}$ の利得の順序となり，囚人のジレンマゲームと同様，相手の行動に係わらず D を選択した方が自分の利得が大きくなる．ただし，囚人のジレンマゲームと異なり，両者がより大きな利得を得られる行動の組合せは他にないため，両者は自分の利得を追求すべきである．

このゲームの場合，効用利用 Q 学習では両者が C を選択する，つまりあえて利得を追求しない行動を選択するようになってしまうが，学習率調整 Q 学習では，学習率の範囲が 0 から 1 であることを利用して，この場合に C を選択するような学習が起こらないことが証明された．

効用利用 Q 学習と学習率調整 Q 学習のシミュレーション実験の結果の例を以下に示す．まず囚人のジレンマゲームにおける結果である．一番上が通常の Q 学習，2 番目が効用利用 Q 学習，一番下が学習率調整 Q 学習の結果である．

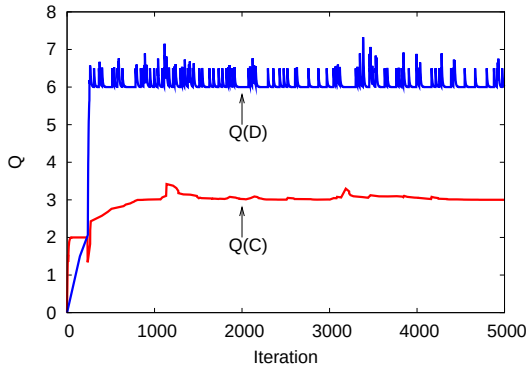


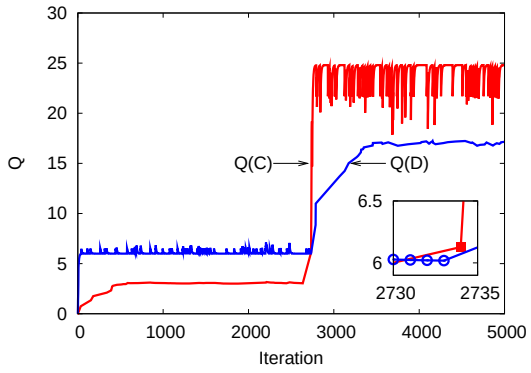


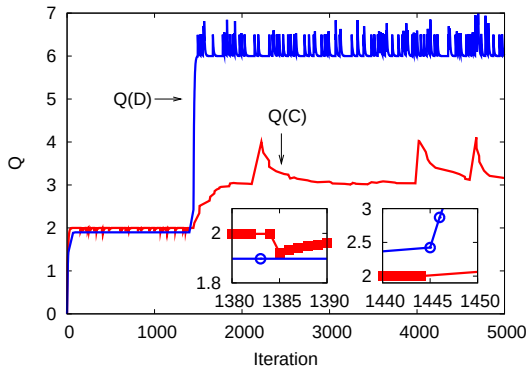


通常の Q 学習では常に D の Q 値（Q(D)）が C の Q 値（Q(C)）よりも大きく，エージェントは D を選択してしまうが，提案手法はどちらもある時点より後では Q(C) ＞ Q(D)となっており，C を選択するようになったことが分かる．この逆転する時刻は現在のところランダムである．

次に，行き詰まりゲームでのシミュレーション実験結果の例を示す．同様に一番上が通常の Q 学習，2 番目が効用利用 Q 学習，一番下が学習率調整 Q 学習の結果である．







通常のQ学習では望ましい行動であるDを選択することが分かるが，効用利用 Q 学習では望ましくないCを選択するようになってしまうことが分かる．一方で，学習率調整 Q 学習では望ましい行動 D を選択することが分かる．

このような学習手法の提案のほかに，2 人 2 行動同時手番対称ゲームの枠組みからは外れるが，他者に関する知識を持たずに行動を選択する例として，株式などの市場取引における通常の強化学習の適用について調査を行い，結果を公表した．

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 10 件）

① Koichi Moriyama. "Learning-Rate Adjusting Q-learning for Two-Person Two-Action Symmetric Games". *Proceedings of the third KES Symposium on Agent and Multi-Agent Systems - Technologies and Applications, KES-AMSTA 2009 (Lecture Notes in Artificial Intelligence 5559),* pp. 223-232, 2009 (in press). 査読有

② Koichi Moriyama. "Learning-Rate Adjusting Q-learning for Prisoner's Dilemma Games". *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT'08,* pp. 322-325, 2008. 査読有

③ 森山 甲一. "2 人 2 行動ゲームのための学習率調整Q学習". 合同エージェントワークショップ＆シンポジウム 2008 (JAWS2008) 論文集, 2008. 査読有

④ 森山 甲一. "囚人のジレンマゲームにおけるQ学習による協調の維持". コンピュータソフトウェア, Vol. 25, No. 4, pp. 145-153, 2008. 査読有

⑤ Koichi Moriyama, Mitsuhiro Matsumoto, Ken-ichi Fukui, Satoshi Kurihara, and Masayuki Numao. "Reinforcement Learning on a Futures Market Simulator". *Journal of Universal Computer Science,* Vol. 14, No. 7, pp. 1136-1153, 2008. 査読有

⑥ Koichi Moriyama. "Utility Based Q-learning to Maintain Cooperation in Prisoner's Dilemma Games". *Proceedings of the 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology, IAT'07*, pp. 146-152, 2007. 査読有

⑦ 森山 甲一. "囚人のジレンマゲームにおけるQ学習による協調の維持". 合同エージェントワークショップ＆シンポジウム 2007 (JAWS2007) 論文集, 2007. 査読有

⑧ 森山 甲一. "囚人のジレンマゲームにおけるQ学習による協調の維持". 第 6 回情報科学技術フォーラム (FIT2007) 講演論文集, pp. 419-422, 2007. 査読無

⑨ Koichi Moriyama, Mitsuhiro Matsumoto, Ken-ichi Fukui, Satoshi Kurihara, and Masayuki Numao. "Reinforcement Learning on a Futures Market Simulator".

Proceedings of the first KES Symposium on Agent and Multi-Agent Systems - Technologies and Applications, KES-AMSTA 2007 (Lecture Notes in Artificial Intelligence 4496), pp. 42-52, 2007. 査読有

⑩ 松本 光弘, 福井 健一, 森山 甲一, 栗原 聡, 沼尾 正行. "U-MartにおけるQ学習エージェントの設計と評価", 人工知能学会全国大会（第 20 回）論文集, 1B2-2, 2006. 査読無

６．研究組織

(1) 研究代表者

森山　甲一　(MORIYAMA KOICHI)
大阪大学・産業科学研究所・助教
研究者番号：10361776

(2) 研究分担者

なし

(3) 連携研究者

なし