|  |  |
|---|---|
| Title | 不確実な状況における利己的な学習主体の相互協調 |
| Author(s) | 鳥居，拓馬 |
| Citation |  |
| Issue Date | 2016-03 |
| Type | Thesis or Dissertation |
| Text version | ETD |
| URL | http://hdl.handle.net/10119/13511 |
| Rights |  |
| Description | Supervisor:橋本　敬，知識科学研究科，博士 |

| | | | | |
|---|---|---|---|---|
| 氏名 | 鳥居拓馬 | | | |
| 学位の種類 | 博士(知識科学) | | | |
| 学位記番号 | 博知第 182 号 | | | |
| 学位授与年月日 | 平成 28 年 3 月 24 日 | | | |
| 論文題目 | 不確実な状況における利己的な学習主体の相互協調 | | | |
| 論文審査委員 | 主査 | 橋本　　敬 | 北陸先端科学技術大学院大学 | 教授 |
| | | Huynh Nam Van | 同 | 准教授 |
| | | Dam Hieu Chi | 同 | 准教授 |
| | | 中森　義輝 | 同 | 客員教授 |
| | | 野田　五十樹 | 産業技術総合研究所 | 総括研究主幹 |

## 論文の内容の要旨

Cooperation is a common form of social interaction. The problem of cooperation is a conflict between individual rationality and collective rationality: selfish players maximizing their own profit obtain a smallest profit as a collective. Since in many situations in nature or society behaving selfishly is apparently a profitable option, how organisms solve the problem of cooperation is a long-lasting question. The problem was formulated as Prisoner's Dilemma in game theory, and various theoretical and empirical studies have shown possibilities of cooperation.

In game theory, a rational player takes a payoff-maximizing strategy (or action). A payoff to each player is a function of strategies of all players. A profile of strategies is called a Nash equilibrium if no one can improve his payoff by unilaterally changing his strategy. Prisoner's Dilemma is a game, in which each of two players has two options: Cooperate or Defect. It has been proved that the only Nash equilibrium in one-shot Prisoner's Dilemma is mutual defection. Iterated Prisoner's Dilemma (IPD) is its extensive-form variant. It has been proved that mutual cooperation can be a Nash equilibrium as well as mutual defection if both players consider sufficiently long-term future. One of the prominent findings in this area is the so-called tit-for-tat (TFT) or reciprocal strategy, whose behavioral rule is: if you cooperate (defect) I will cooperate (defect). Many theoretical studies has provided evidence supporting TFT for cooperation in some context, however, for what objective one acquires TFT-like behaviors is an unanswered question.

A key idea in recent game theory is bounded rationality: a decision maker has to choose an action based on limited information and restricted cognitive resources. Learning is a means to overcome uncertainty arising from incompleteness of information. Some psychological studies have shown that human cooperation is observed more frequency under uncertainty. Chapter 1 and 2 contains the background and reviews regarding the problem of cooperation in psychology and game theory.

This thesis aims at proposing a theoretical description for the problem of cooperation between two learning players under uncertainty. Concretely, the thesis aims at showing some conditions for mutual cooperation and what the learning players can maximize to establish mutual cooperation. Led by findings from psychology and game theory, in this thesis, the problem is formulated as Iterated Prisoner's Dilemma under uncertainty, where one can only get feedbacks to him in response to his actions. That is, no one can get information about the payoff matrix and related to the opponent. Reinforcement learning is a mechanism that can maximize its total profit only based on feedbacks (payoffs) in response to its actions. There are many evidence that reinforcement learning can solve various real world problems in uncertain environments.

In this thesis, we showed some conditions for mutual cooperation that can be established between selfish, reinforcement learners who attempt to maximize their own profit under uncertainty. Mutual cooperation is observed almost surely if both players make decisions based on sufficiently long-term experience. From a detailed analysis, TFT-like behaviors are observed during mutual cooperation between reinforcement learners. This finding gives TFT a position as a by-product of selfish/greedy learning. Chapter 3 includes the findings above.

Further conditions are investigated by approximating IPD of reinforcement learners. Using this approximated model, we formally studied several properties of the game, including payoff-related conditions for mutual cooperation. Based on the findings from the approximated model, we derived a payoff matrix that dramatically improve mutual cooperation between reinforcement learners. This can be interpreted as a practical mechanism for cooperation. The approximated model was studied in Chapter 4.

Chapter 5 contains a framework for future studies, which allows us to study reinforcement learning strategies within a more generalized class, although there are technical issues. In this thesis, we studied 1st-order strategy class, and discussed future directions.

Combining all the findings in this thesis and previous studies, a theoretical description for the problem of cooperation between learning players is argued in Chapter 6. It states that mutual cooperation can be established under uncertainty if both selfish players learn to maximize their own profit from their long-term trial-and-error experience.

The problem of cooperation is a special case of the free-rider problem and/or shared resource problem, which are more common in real social situations. For example, one reported that leaving the problem of cooperation unsolved declines the performance of team members. The problem of cooperation can be one of the common barriers that impede performance of organizational activity, such as organizational knowledge creation. The findings in this thesis will provide a clue to resolve conflict situations in real social interactions. Contrary to a common belief about uncertainty, our finding suggests the possibility that uncertainty regarding information, especially conflict relationship, might improve mutual cooperation, if participants can learn from their action-feedback experience.

**Keywords:** Cooperation, Reinforcement learning, Game theory, Uncertainty, Prisoner's dilemma

## 論文審査の結果の要旨

複数の個体が協調的に振る舞う状態がいかにして実現されるかは自然界や社会で重要な問題であり，ゲーム理論や進化ゲーム理論を用いて，経済学，社会学，組織論，政治学，心理学，進化生物学といった人間と社会に関わる様々な分野で長く興味を持たれ現在も研究が続けられている. 人間が生来持つと考えられる協調的な性質がいかに獲得されたかを問うには進化を考えることが重要であるが, 人々が相互作用して組織を作る社会の現実的な状況において協調状態を実現する方法を考える基礎としては, 学習による到達条件やメカニズムを探ることがより有用であると考えられる．そこでこの研究では，人間の認知能力の限界（限定合理性）を重視し，直面している利害関係や相互作用する他者に関する情報が非常に限られている状況（本論ではこれを「不確実性の高い問題状況」としている）において，学習によりいかにして協調状態に到達することができるかを探求している．具体的には，自身の行動とそれに対して得た利得のみを知り得る強化学習プレイヤが，繰り返し囚人のジレンマ（IPD）ゲームをプレイする中で，互いに協調的行動を取る相互協調状態にいたる条件を，数理的および数値的な解析により求めた.

本論文では，強化学習の記憶パラメータが高い領域，すなわち，過去の自身の行動と利得の履歴を十分考慮して自己の利得を最大化するように意思決定をする場合に, 強化学習戦略をとるプレイヤ同士で相互協調が実現されることを明らかにした. 情報が限られている状況が協調状態に導くという逆接的な結果は, 情報の扱いが重要なゲーム理論の研究の中で新規性のある知見である．そして，記憶パラメータに対する行動パターンを分析し，記憶が短いときは相互裏切り的戦略を，記憶が長くなるにつれしっぺ返し的な戦略を学習し，それにより相互協調が実現されることを示した. しっぺ返し戦略をアプリオリに与えずいかにして学習により獲得されるかを示すこの結果も新規性と学術的な価値がある．さらに，強化学習戦略による IPD ゲームを近似的に表現した数理モデルのベクトル場を分岐解析するという新しい解構造の分析方法を提案し，「相互協調の利得が正, 相互裏切りの利得が負であり, それぞれ協調裏切りの組の利得との差が小さい」という相互協調解の存在条件を示した．これは直観的に分かりやすい結果であり，このような条件がなぜ成り立つかを数理的にきちんと示すことは, この条件をどう一般化できるかを論じられるため学術的価値が高い．実際に，公共財ゲーム等より社会的な状況へ拡張できることも示している.

以上，本論文では，強化学習プレイヤによる繰り返し囚人のジレンマゲームにおいて，限られた情報の下で履歴を十分考慮して自己利益を最大化する強化学習により相互協調が実現できることとその条件を示した．この研究は，組織において相互協調状態を実現するには，どのような報酬構造と情報環境を作ればいいかに指針を与えるという有用性もある．このように，本研究

は知識科学への学術的貢献が大きい．よって博士（知識科学）の学位論文として十分価値あるものと認めた．