様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　　元 年　　６ 月 ２５ 日現在

機関番号：５４６０１
研究種目：基盤研究(C)（一般）
研究期間：2016～2018
課題番号：１６Ｋ００３１７

研究課題名（和文）継続的強化学習エージェントとコーチ役による自律学習システムの設計

研究課題名（英文）Designing the autonomous learning system by the continuous reinforcement learning agent with the coach

研究代表者

山口　智浩（Yamaguchi, Tomohiro）

奈良工業高等専門学校・情報工学科・教授

研究者番号：００２４０８３８

交付決定額（研究期間全体）：（直接経費）　　3,500,000円

研究成果の概要（和文）：本研究では，人の継続的な学習能力の向上およびその上達過程の可視化に向けて，人が理解しやすい学習過程の可視化機能を持つ継続的な自律学習システムを構築した．学習者が発見した冗長解が持つ派生目標を学習目標空間上で提示し，派生目標間の位置関係を可視化することで，未発見目標領域（空白域）を間接的に可視化する手法を実現した．被験者による比較実験の結果，学習者の発見した目標が既存目標領域に近いか遠いか，すなわち学習の空白域との関係を示唆する提示条件が上達過程において学習フィードバック情報として重要で，未知の価値観への気づきを促す条件であることが示唆された．

研究成果の学術的意義や社会的意義
近年，注目されている深層学習の主な弱点は(1)人が実現不能な学習手法と(2)内部の学習過程の理解困難さである．これに対し，本研究では深層学習の弱点を補うため，(1)様々な問題を生成し提供することで，人が学習の仕方を学べる機能，(2)学習結果の解釈を行い，人が理解しやすくなるように学習過程・上達過程を可視化する機能を考案した．
本研究によって学習目標となる報酬設計が難しかった強化学習法の幅広い分野への適用が可能になる．また，自律学習システムは問題領域ごとに初期問題を与えると様々な派生問題とその解を反復的に生成するため，問題や解のバリエーションを大量に必要とするタスクに応用できる．

研究成果の概要（英文）：This research proposed the autonomously continuous learning system by visualizing the learning processes which is easy to understand for a human. The main objective of this research is the continuous improvement for a human learning skill and the visualization of the improvement process. In this research, we investigated the method for indirectly visualizing the gap area of undiscovered goals by visualizing the positional relation among derived goals in an automated way. The result of the comparative experiment by the human subjects suggested that it is important that the display condition which indicates the positional relation with the gap of learning (which is the distance between a new goal found by the learner and the area of known goals) as the learning feedback information during the improvement process. In other words, it is the condition to facilitate the awareness of the learner’s unknown sense of values.

研究分野：強化学習

キーワード：機械学習　学習過程　自律学習　逆強化学習　継続的学習　多目的強化学習　目標生成　報酬生起確率

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

本研究で用いる強化学習法は，学習目標を環境中の報酬として設定することで，学習者が試行錯誤を通して自律的に目標への行動系列を学習する手法で，ロボットやエージェントの行動学習として広く用いられている．強化学習法の基本的課題のひとつとして，学習性能を大きく左右する学習目標の与え方(報酬設計法の確立)があり，最終的な研究目標は自律学習のための学習目標の自動生成である．この課題に対する最近の研究としては，入力された行動系列を最適解とする学習目標を求める逆強化学習(例えば，人物行動予測への応用)，最適解の学習を高速化するための副目標の自動生成，環境や目標の変化に対する多目的強化学習，学習目標に対する最適解と最適条件を一括に収集する一括強化学習等が提案されている．しかしながら，これら既存研究では，学習の収束を目的とした研究が中心であり，自律学習のための目標の自動生成研究は，学習目標を正当化するメタ知識の獲得問題が生じるため，申請者の知る限り行われていない．

こういった国内外の研究動向に対して，本研究ではビジネス心理学分野で提案された継続的改善を目的とした人の学習モデルに着目し，自ら学習目標を生成する自律学習エージェントを設計する．具体的には，表１に示す人とエージェントの学習モデルの比較からその差異を埋めることに取り組む．表 1(a)に示す心理学分野での人の継続的学習モデル[5]と表 1(b) に示す人工知能分野での強化学習モデルから分かるように，表 1(a)では継続的改善のため学習目標への“気づき”と“振り返り”があるのに対して，表 1(b)ではステップ１と５の学習前/学習後プロセスがなく，学習エージェント自身で学習目標を継続的に発見・生成するメカニズムが存在しない．そこで，本研究では“気づき”と“振り返り”に相当する機構を考案してエージェントに導入することによって，エージェントが人を適切に学習させるための目標(＝新たな問題)を生成し，人の継続的な学習を促す継続的な自律学習システムの実現を目指す．

表1　人とエージェントの学習モデル比較

	(a) 継続的学習	(b) 強化学習
1	気づき	(学習前プロセス)
2	理解	モデル同定
3	目標の公約	与えられた学習目標
4	行動化	最適解の探索
5	振り返り	(学習後プロセス)

２．研究の目的

本研究では，人の継続的な学習能力の向上およびその上達過程の可視化に向けて，(1)人が「一を聞いて十を知る」ために冗長解の収集と冗長解からの創造的な問題の生成を繰り返して学習させるエージェントと，(2)人が理解しやすい学習過程の可視化手法を探究するとともに，それを統合した継続的な自律学習システムを構築し，その有効性を検証することを目的とする．具体的な研究目標は下記の３点である．

(1)学習者を適切に学習させるための目標生成の自動化
(2)学習者の上達過程を記述・説明する学習目標空間の可視化
(3)学習者とのやり取りを通して学習させる継続的強化学習エージェントの設計

３．研究の方法

(1)学習者を適切に学習させるための目標生成の自動化

学習者の上達を支えるコーチ役(学習者を適切に学習させるための目標(＝問題)を生成する)エージェントを設計するために，冗長解が潜在的に持つ未知の価値観への気づきを促す機能を探究するとともに，冗長解を振り返り，新たな学習目標を追加することで派生問題を生成する機能を考案する．

(2)学習者の上達過程を記述・説明する学習目標空間の可視化

学習者の上達過程を記述・説明するために，解の規模を表す“解の長さ”と，学習目標の分かりにくさを表す“獲得報酬の情報量(生起確率の逆数)の和”を軸として学習目標空間を定義する．例えば，既存の学習結果に対し，解が長くて分かりにくさが大きい学習目標を持つ問題を解ければ，より難しい問題へと学習者が上達したと言える．この変化によって上達過程を表現する．

(3)学習者とのやり取りを通して学習させる継続的強化学習エージェントの設計

獲得報酬の生起確率に基づく一括強化学習手法と一括逆強化学習手法を考案する．本研究では冗長解集合から学習用の様々な問題を生成し，それぞれの学習効果を学習目標(＝問題)ごとに定量化するため，冗長解集合を一括で得る強化学習と，各冗長解から学習目標を作成する一括逆強化学習が重要となる．そこで，本研究では報酬獲得解(＝学習結果)に対し各報酬の生起確率を生起確率ベクトルとして算出し，解を高速かつ網羅的に収集する強化学習手法を基にする．

４．研究成果

(1)学習者を適切に学習させるための目標生成の自動化

学習者の上達を支えるコーチ役(学習者を適切に学習させるための目標(＝問題)を生成する)エージェントを設計するために，冗長解が潜在的に持つ未知の価値観への気づきを促す機能として，まず，冗長解を振り返り，冗長解上に新たな学習目標を追加することで派生問題を生成する機能を考案した．次に冗長解の潜在的な価値推定の自動化手法として冗長解の逆強化学習によって新たな目標の推定を実現した.派生的な要素技術として,深層学習で得られたルールを,学習目標の起点となる初期ルールとする活用について検討した．これらの研究成果を主な発表論文として，雑誌論文[3,4,6,7,9]，学会発表[2,4,5,6,8,10,11,12]，図書[3]等で発表した.

(2)学習者の上達過程を記述・説明する学習目標空間の可視化
学習目標空間での冗長解と派生目標との関係を可視化するために，初年度作成した実験システムを用いて，学習者が発見した冗長解についてその解が持つ派生目標を学習目標空間上で提示し，発見した派生目標間の位置関係を可視化することで，未発見目標領域（空白域）を間接的に可視化する手法を実現した．その有効性を実験的に検証するために被験者による比較実験を行なった．その結果，学習者の発見した目標が既存目標領域に近いか遠いか，すなわち学習の空白域との関係を示唆する提示条件が上達過程において学習フィードバック情報として重要で,未知の価値観への気づきを促す条件であることが示唆された．これらの研究成果を主な発表論文として，雑誌論文[4,5,6,8]，学会発表[12,13]，図書[2,3]等で発表した.

(3)学習者とのやり取りを通して学習させる継続的強化学習エージェントの設計
獲得報酬の生起確率に基づく一括強化学習手法として，報酬獲得解(＝学習結果)に対し各報酬の生起確率を生起確率ベクトルとして算出し，解を高速かつ網羅的に収集する強化学習手法を基にして，与えられた初期問題の最適解や冗長解を網羅した報酬獲得解集合を求める一括強化学習手法を考案した．さらに，報酬獲得解が生起確率ベクトル空間の点に対応することから,解集合となる点集合から凸包の各頂点を算出して，多数の冗長解から高速かつ網羅的に凸包の頂点となる代表的な学習目標の生成手法を考案した．これらの研究成果を主な発表論文として,雑誌論文[1,2,3]，学会発表[1,3,7,9,14,15]，図書[1]等で発表した.

５．主な発表論文等

〔雑誌論文〕（計９件）

1. Yamaguchi, T., Nagahama, S., Ichikawa, Y. and Takadama, K.: “Model-based Multi-Objective Reinforcement Learning with Unknown Weights”, Human Interface and the Management of Information, Lecture Notes in Computer Science, 査読有, Springer-Verlag, 2019年(7/29-31), [DOI] to appear　印刷中

2. Uwano, F. and Takadama, K.: “Strategy for Learning Cooperative Behavior with Local Information for Multi-agent Systems”, Principles and Practice of Multi-Agent Systems, Lecture Notes in Computer Science, 査読有, Vol.11224, Springer-Verlag, pp.663-670, 2018年(11/1), [DOI] https://doi.org/10.1007/978-3-030-03098-8_54

3. Uwano, F., Tatebe, N., Nakata, M., Tajima, Y., Kovacs, T. and Takadama, K.: “Multi-Agent Cooperation Based on Reinforcement Learning with Internal Reward in Maze Problem”, SICE Journal of Control, Measurement and System Integration (JCMSI), 査読有,Vol.11, No.4, pp.321-330, 2018年 [DOI]https://doi.org/10.1007/978-3-319-92043-6_52

4. Okudo, T., Yamaguchi, T. and Takadama, K.: “Generating Learning Environments Derived from Found Solutions by Adding Sub-goals toward the Creative Learning Support”, Human Interface and the Management of Information, Lecture Notes in Computer Science, 査読有, Vol.10905, Springer-Verlag, pp.313-330, 2018年(7/19)
[DOI] https://doi.org/10.1007/978-3-319-92046-7_28

5. Yamaguchi, T., Tamai, Y., Honma, Y. and Takadama, K.: Analyzing the Goal Finding Process of Human's Continuous Learning with the Reflection Subtask”, SICE Journal of Control, Measurement, and System Integration, 査読有, Vol.11, No.1, pp.40-47, 2018年(3/6), URL: 10.9746/jcmsi.11.40

6. Okudo, T., Yamaguchi, T., Murata, A, Tatsumi, T., Uwano, F. and Takadama, K.: “Supporting the Exploration of the Learning Goals for a Continuous Learner　Toward Creative Learning”, Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), 査読有, Vol.21, No.5, pp. 907-916, 2017年(9/20), URL: 10.20965/jaciii.2017.p0907

7. Matsumoto, K., Tatsumi, T., Sato, H., Kovacs, T. and Takadama, K.: ”XCSR Learning from

Compressed Data Acquired by Deep Neural Network”, Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), 査読有, Vol. 21, No. 5, pp. 856-867, 2017年(9/20), URL: 10.20965/jaciii.2017.p0868

8. Okudo, T., Takadama, K. and Yamaguchi, T.: ”Designing the learning goal space for human toward acquiring a creative learning skill”, Human Interface and the Management of Information, Part II, Lecture Notes in Computer Science, Vol.10274, Springer-Verlag, 査読有, pp. 62-73, 2017年(7/12)

9. Uwano, F., Tatebe, N., Nakata, M., Takadama, K. and Kovacs, T.: “Reinforcement Learning with Internal Reward for Multi-Agent Cooperation: A Theoretical Approach”, EAI Endorsed Transactions on Collaborative Computing, 査読有, Vol.16, Issue 8, pp.1-8, 2016年, http://dx.doi.org/10.4108/eai.3-12-2015.2262878

〔学会発表〕(計１５件)

1. Takadama, K., Yamazaki, D, Nakata, M. and Sato, H.: “Complex-Valued-based Learning Classifier System for POMDP Environments”, 2019 IEEE Congress on Evolutionary Computation (CEC2019), 2019年(6/12)

2. Hasegawa, S., Uwano, F. and Takadama, K.: “Maximum Entropy Inverse Reinforcement Learning with incomplete expert”, The 24th International Symposium on Artificial Life and Robotics (AROB 2019), pp.361-366, 2019年(1/23)

3. 上野 史, 高玉 圭樹: ”報酬の動的変化に適応する通信なしマルチエージェント協調学習のための公平性に基づく内部報酬設定法”, 計測自動制御学会, システム・情報部門 学術講演会2018 (SSI2018), 2018年(11/25)

4. 長谷川 智, 上野 史, 高玉 圭樹: ”行動系列分割に基づく不完全なエキスパートからの逆強化学習”, 計測自動制御学会, システム・情報部門 学術講演会 2018 (SSI2018), 2018年(11/25)

5. Matsumoto, K., Takano, R., Tatsumi, T., Sato, H., Kovacs, T. and Takadama, K.: ”XCSR Based on Compressed Input by Deep Neural Network for High Dimensional Data”, International Workshop on Learning Classifier Systems (IWLCS 2018) in Genetic and Evolutionary Computation Conference (GECCO 2018), 2018年(7/15-16)

6. 長谷川 智, 梅内 祐太, 上野 史, 佐藤 寛之, 山口 智浩, 高玉 圭樹: ”負の報酬生成による環境変化に適応可能な逆強化学習”, 計測自動制御学会, 第45回知能システムシンポジウム, 2018年(3/8)

7. 長濵 将太, 市川 嘉裕, 高玉 圭樹, 山口 智浩: ”報酬生起確率ベクトルに基づくあらゆる状況に対する強化学習”, 計測自動制御学会, 第45回知能システムシンポジウム, 2018年(3/8)

8. 松本 和馬, 高野 諒, 上野 史, 佐藤 寛之, 高玉 圭樹: ”深層学習による次元圧縮ルールの学習分類子システムにおける初期ルールとしての可能性”, 進化計算学会, 第11回進化計算シンポジウム 2017, pp.168-175, 2017年(12/9)

9. 長濵 将太, 山口 智浩, 高玉 圭樹: ”報酬生起確率ベクトルと重みベクトルに基づく全ての最適方策の一括強化学習”, 計測自動制御学会, システム・情報部門 学術講演会 2017 (SSI2017),
SS13-2, pp.832-836, 2017年(11/25)

10. 松本 和馬, 高野 諒, 佐藤 寛之, 高玉 圭樹: ”深層学習による圧縮ルールを復元する学習分類子システムとその精度向上”, 第13回進化計算学会研究会, 進化計算学会, pp.98-101, 2017年(9/2)

11. 福田 千賀, 村田 暁紀, 石井 晴之, 佐藤 寛之, 高玉 圭樹: “難易度と技術偏差に基づく学習目標生成を促すインタラクティブ学習支援”, 計測自動制御学会, 第44回知能システムシンポジウム, 2017年(3/14)

12. Okudo, T., Yamaguchi, T. and Takadama, K.: “Designing the learning goal space toward acquiring a creative learning skill”, The 22nd International Symposium on Artificial Life and Robotics (AROB'17), 2017年(1/21)

13．玉井 雄貴，山口 智浩，高玉 圭樹:“サブゴールの振り返りによる学習者の継続的学習支援”，計測自動制御学会，システム・情報部門 学術講演会 2016（SSI2016），2016年(12/7)

14．Uwano，F．and Takadama，K.:“Communication-less Cooperative Q-learning Agents in Maze Problem”，The 20th International Symposium on Intelligent and Evolutionary Systems (IES 2016)，Intelligent and Evolutionary System，Springer-Verlag，pp.453-467，2016年(11/17)

15．Saito，R.，Nakata，M.，Sato，H.，Kovacs，T．and Takadama，K.:”Preventing Incorrect Opinion Sharing with Weighted Relationship among Agents”，The 18th International Conference on Human-Computer Interaction (HCI International 2016)，Lecture Notes in Computer Science，Vol.9735， Springer-Verlag，pp.50-62，2016年(7/20)

〔図書〕（計３件）

1．著者名：Dan Zhang and Bin Wei (Eds.)，Yamaguchi，T.，Nishimura，T.，Nagahama，S．and Takadama，K.
出版社名：IGI Global
書名：Novel Design and Applications of Robotics Technologies，Chapter 9,"Awareness Based Recommendation by Passively Interactive Learning: Toward a Probabilistic Event"
発行年：2018年，総ページ数：341(pp.247-275)

2．著者名：Maki Habib (Eds.)，Yamaguchi，T.，Tamai，Y．and Takadama，K.
出版社名：IGI Global
書名:Handbook of Research on Biomimetics and Biomedical Robotics，Chapter 19，“Analyzing the goal finding process of human's learning with the reflection subtask”
発行年：2017年，総ページ数：532（pp.）

3．著者名：Maki Habib (Eds.)，Okudo，T.，Yamaguchi，T. and Takadama，K.
出版社名：IGI Global
書名:Handbook of Research on Biomimetics and Biomedical Robotics，Chapter 20，“Designing the Learning Goal Space for Human Toward Acquiring A Creative Learning Skill”
発行年：2017年
総ページ数：532（pp.）

〔その他〕
なし

６．研究組織
(1)研究分担者
高玉　圭樹（TAKADAMA，Keiki）
電気通信大学・大学院情報理工学研究科・教授
研究者番号：20345367

(2)研究協力者
なし