

CBR によるマルチエージェント強化学習における 状態空間の削減について

1P-02

藤巻伸洋 大園忠親 新谷虎松

名古屋工業大学 知能情報システム学科

e-mail: {fujimaki, ozono, tora}@ics.nitech.ac.jp

1 はじめに

本研究は、マルチエージェント強化学習における状態空間の爆発問題を、事例ベース推論を利用し、解決することを目的とする。マルチエージェント強化学習とは、複数のエージェントが強化学習によって、問題の解決法を学習する手法である。しかし、強化学習では、認識する対象の増加により、エージェントが認識すべき状態が指数関数的に増加する、状態空間の爆発問題という問題がある [2]。本研究では、「類似する問題は類似する解を持つ」ことを利用した問題解決手法である事例ベース推論の枠組みを利用し、状態空間の爆発問題を解決する。

2 強化学習における状態空間爆発問題

強化学習エージェントは、環境の状態を入力とし、それに対する行動を出力として返す。その行動によって環境は変化し、報酬を発生させる。発生した報酬によってエージェントは、行動を評価し、次の行動決定に利用する。このサイクルの繰り返しにより学習を進めていく。強化学習では、学習対象が複雑になるにつれ、状態数が増加する。状態数の増加は、状態の検索速度に影響を及ぼす。また、同じ状態に遭遇する確率が低下し、試行錯誤によって学習する強化学習では、状態に適した行動を判断することが困難なものとなる。特に、エージェントが認識することのできる空間、対象が増加すると、状態数は指数的に増加する。これを強化学習における状態空間の爆発問題と呼ぶ。状態空間の爆発問題に対して、(1) 状態表現を連続化し、政策を関数近似表現する方法、(2) 階層表現の利用、(3) 他の機械学習法の導入、の三種類の手法が提案されている [2]。本論文では三番目の手法に基づき、強化学習に事例ベース推論 (Case-Based Reasoning: CBR)[3] の枠組みを取り入れることで、状態空間の爆発問題の解決を試みる。

CBR は、「類似した問題の解は類似している」という人の経験則的な推論過程を再現するものである。CBR

On the reduction of the state space in multiagent reinforcement learning based on CBR

Nobuhiro FUJIMAKI, Tadachika OZONO, Toramatsu SHINTANI

Dept. of Intelligence and Computer Science, Nagoya Institute of Technology, Gokiso, Showa-ku, Nagoya 466-8555 JAPAN

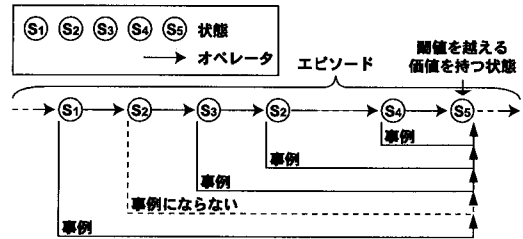


図 1: エピソードからの事例獲得

の特徴は、問題解決のためにルールを抽出しなくとも、過去の問題解決事例の格納によって、知識を蓄えることができる点が挙げられる。今回は、強化学習エージェントによって解決された事例を利用し、過去の事例と類似する状態において、過去の事例で実行された行動列を再現するために、CBR の枠組みを利用する。

3 強化学習に対する CBR の利用

強化学習エージェントは、無限回の試行によってすべての状態を経験し、それぞれの状態での最適な行動を学習するように行動する。しかし、実際の問題に適用することを考えると、無限回の試行や、すべての状態を体験するには、膨大なエピソードを蓄積する必要がある。エピソードとは、問題の初期状態からゴール状態までの状態行動対の列である。ここで、類似するエピソードを事例としてまとめることで、使用するメモリ量を抑えることが可能である。事例とは、始点、および終点となる二つの状態と、それらをつなぐ行動の列である。事例は、エピソードを基に作成される。エピソードからの事例獲得の様子を図 1 に示す。エピソードから閾値を超える価値を持つ状態を抽出する (図 1 の S_5)。状態の価値は $V(s) = \max_a Q(s, a)$ で表される。ここで、 s は状態、 a は行動、 $Q(s, a)$ は、状態 s 、行動 a での Q 値である。抽出した状態行動対の状態を事例の終点とする。そして終点以前の状態全てを始点として (図 1 の S_1, S_2, S_3, S_4)、その間の行動の列を事例の行動列とする。同じ始点で、かつ、同じ終点の事例がある場合は、より行動列の短い事例を残す (図 1 では S_2 から始まる二つの事例がこれに相当する)。

新たな事例を獲得すると、事例を利用する行動のQ値を設定する。事例の始点における事例利用のQ値は、事例の終点の価値とする。これは、その事例を利用すると終点の状態にまで遷移していく可能性があることから、事例利用のQ値として利用している。

エージェントによる事例利用は以下のような処理を行う。エージェントは、獲得された事例の中から、現在認識している状態ともっとも類似する状態を始点とする事例を選択する。状態の類似度は、式(1)を用いて計算する。

$$\text{sim}(s, t) = \sum_i^N \sum_j^N d_{ij} \quad (1)$$

$$d_{ij}(s, t) = \begin{cases} N - (|i - N/2| + |j - N/2|) & (s_{ij} = t_{ij}) \\ 0 & (s_{ij} \neq t_{ij}) \end{cases} \quad (2)$$

エージェントが認識する状態は、エージェントを中心とした一辺の長さNの正方形である。 s, t はともに状態、 s_{ij}, t_{ij} は、状態 s, t の座標 i, j を示す。式(2)によって、エージェントにより近い座標に存在するものに対して重みを与え、類似度に反映している。エージェントは、検索された事例で実行された行動列を一つずつ再現していく。

本手法を倉庫管理問題に適用する。倉庫管理問題とは、Grid上に配置された荷物を、指定の目的地までエージェントによって運ぶ問題である。エージェントが取ることでできる行動は「前進」「右90度回転」「左90度回転」「事例利用」の四種類。エージェントが荷物のあるGridに前進した場合、さらにその先に荷物がなければ、エージェントは荷物を押すことができる。押した先が目的地だった場合、荷物は片づけられ、エージェントは環境から報酬を与えられる。環境内のすべての荷物が目的地に片づけられた時点で一つの問題は終了する。エージェントの学習にはQ-learning、行動の選択には $\epsilon = 0.1$ の ϵ グリーディを用いた[1]。

4 評価

本手法を適用したエージェントと、適用しないエージェントで同じ問題を解決させ、どちらがどれだけの状態を学習し、保持しているかを比較する。

5×5のGridに二体のエージェント、二つの荷物、一つの目的地を配置し、荷物を目的地に運び込んだエージェントは報酬を獲得する。二つの荷物がなくなった時点で一つのエピソードは終了とする。

図2には、それぞれのエージェントが学習した状態数の差を示した。横軸にエピソード、縦軸にそれまで

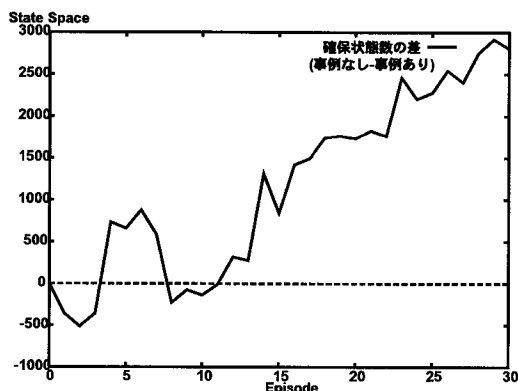


図2: 事例利用の有無による保持状態数の差

に経験した(エージェントがメモリに保持している)状態数の差を示している。ここでの差とは、(事例を使わないエージェントの状態数)-(事例を使ったエージェントの状態数)である。実験開始直後では、事例が蓄積されていないため、状態数はどちらが少ないとは断定できない。しかし、エピソード数10を越えた辺りで状態数の差が広がり、事例を使ったエージェントの方がより少ない状態数で問題を解決している。これは利用可能な事例が蓄積され、それを用いて問題を解決しているためと考えられる。よって、未知の状態にまで探索を広げなくとも、問題の解決が可能であることが示されていると言える。

5 おわりに

本論文では、マルチエージェント強化学習において事例ベース推論を利用することで、状態空間の爆発問題を解決することを目指した。その手法として、エピソードからの事例獲得方法と、状態の類似度の計算手法を示し、実験によって状態数を抑えることが可能であることを示した。

参考文献

- [1] Richard S. Sutton, Andrew G. Barto: "Reinforcement Learning-An Introduction-", The MIT Press, (1998).
- [2] 荒井幸代: "マルチエージェント強化学習-実用化に向けての課題・理論・諸技術との融合-", 人工知能学会誌, Vol. 16, No. 4, pp.476-481, (2001).
- [3] Ian Watson: "Applying Case Based Reasoning", Morgan Kaufmann Publishers, Inc., (1997).